The Incompleat Egoist

DAVID GAUTHIER

THE TANNER LECTURES ON HUMAN VALUES

Delivered at
Stanford University

May 10, 1983

DAVID GAUTHIER is a native Torontonian now resident in Pittsburgh. He was educated at the University of Toronto, Harvard University, and the University of Oxford. From 1958 to 1980 he was a member of the Department of Philosophy at the University of Toronto, rising from Lecturer through the ranks to Professor and becoming Chairman in 1974. He is now Professor of Philosophy, Senior Fellow in the Center for Philosophy of Science, and Chairman of the Department of Philosophy at the University of Pittsburgh. In 1979 he was elected a Fellow of the Royal Society of Canada.

Professor Gauthier is the author of two books, editor of a third, and is currently completing a fourth, “Morals by Agreement,” which brings to fruition almost twenty years’ work in developing a contractarian theory of morality. He has also written numerous papers, mainly in moral and political theory. The most enduring of his non-philosophical interests is in light rail transit.

I. *WHAT CAN AN EGOIST DO?*

1. "Egoism . . . is the doctrine which holds that we ought each of us to pursue our own greatest happiness as our ultimate end."[1] Thus G. E. Moore, who proceeded to charge this doctrine with "flagrant contradiction."[2] "The egoistic principle," Brian Medlin asserted, "is inconsistent."[3] In levelling these accusations, Moore and Medlin have been representative of a host of philosophers who have found egoism wanting in rationality. But why accuse the egoist? Left to himself, surely he seeks only to do as well for himself as possible, and this intent, if not wholly attractive, seems to fall squarely within the confines of the economist's utility-maximizing conception of practical rationality — hardly, then, what we should expect to find contradictory or inconsistent.[4] Philosophers are blessed with both talent for and desire of finding paradox where other mortals suspect none, yet what, *rationally*, could be at fault with the attempt to do as well for oneself as possible?

As a philosopher, I have up my sleeve what, if not truly paradoxical, should seem unexpectedly puzzling. But the questions I shall raise about egoism come, not from the traditional philosophical repertoire, but rather from the theory of rational choice.

[1] G. E. Moore, *Principia Ethica* (Cambridge: At the University Press, 1903), p. 76.

[2] Ibid., p. 102.

[3] Brian Medlin, "Ultimate Principles and Ethical Egoism," *Australasian Journal of Philosophy* 35 (1957), p. 118.

[4] Note that the egoist's aim is not dictated by the utility-maximizing conception of practical rationality. What is rational, according *to* this conception, is to do as well as possible — to maximize some measure defined over the possible outcomes of one's actions. The characteristics of the measure to be maximized are left largely unspecified by this maximizing requirement. The egoist adds to the idea of doing as well as possible the specification that the measure be self-directed, so that he do as well *for himself* as possible.

More particularly, although the lone egoist will pass rational scrutiny, yet when put with others of his persuasion, in interaction in which each seeks to maximize his own happiness, grounds for challenging the rationality of egoism appear. And these grounds concern, not so much the egoist's concern with his own happiness, but rather his maximizing principle of choice. Something is amiss in our account of practical rationality.

2. Let us then focus briefly on the theory of rational choice. We may first recall a dictum laid down by John Rawls: "The theory of justice is a part, perhaps the most significant part, of the theory of rational choice."[5] I shall interpret this dictum in a quite un-Rawlsian way, and in order to sketch my interpretation I must temporarily set egoism to one side; but before doing this, let us note an immediate connection between Rawls' claim and our concern with the rationality of egoism. If the theory of justice is literally a part of the theory of rational choice, and so much a part that justice proves to be required in rational choice, then it would seem that either justice is compatible with egoism, or that egoism is not compatible with rationality. If the former is implausible, then we may expect the case for the rationality of justice to be linked to the case against egoism. And both will depend on a proper understanding of rational choice.

Before concluding our reflections we shall indeed have proceeded from an argument against the rationality of egoism to an argument linking, not only justice, but morality, with rational choice. But that link comes at the end of a long chain. Here let us reflect on what Rawls said and generalize it to what I believe — that moral theory as a whole is part of choice or decision theory. What I believe, however, is *not* what Rawls believes.

I treat moral principles as principles *for* rational choice. In a very general and important type of interaction, which I shall call *cooperutive,* a rational actor would — I claim — base his or her

[5] John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1972), p. 16

choice among possible actions on a moral principle, provided he or she expected others to do likewise. In section five of this part I shall consider what a principle for choice, or for action, is, and in the second part I shall explain why a rational actor would base choices on principles appropriately characterized as moral.

Rawls treats the principles of justice, not as principles *for* rational choice, but as objects *of* rational choice.[6] This is a very different matter. For Rawls the principles of justice determine the basic structure of society. He asks, what principles, constitutive of society, would a rational individual choose in the "original position," behind a veil of ignorance making him unaware of his identity except as a free and equal person. Rawls identifies the principles so chosen with the principles of justice. This is how he connects the theory of justice with the theory of rational choice.

Note the differences between us. Rawls asks: *what* would rational actors choose behind a veil of ignorance? He answers: they would choose the principles of justice. I ask: *how* would rational actors choose in cooperative interaction? I answer: they would choose on the basis of moral principles. For Rawls the principles of justice constitute the *solation* to a particular problem of rational choice.[7] For me moral principles are used by persons in *solving* certain problems of rational choice. Rawls uses principles of rational choice as tools in developing his theory of justice. I develop moral theory as part of the theory of rational choice — as part of the theory that determines what principles a rational actor would use in choice.

There is a second equally important difference between Rawls' attempt to use rational choice in characterizing justice and my attempt to develop morality as part of rational choice. And this difference bears directly on my concern with egoism. The theory of rational choice examines two significantly different forms of

[6] My account here reflects sections 2–4, of *A Theory of Justice*.

[7] See *A Theory* of *Justice*, section *20*.

agency, parametric and strategic.[8] The parametric actor chooses in an environment that, whether its characteristics be known to him or not, he treats as fixed in relation to his choice. His choice is a response to circumstances that are not, or are not considered to be, responsive to him as choosing. The strategic actor chooses in an environment that is responsive to him as a chooser. He relates his choices to an environment that includes other actors seeking to relate to his choices. Egoism, we shall find, succeeds for parametric choice but fails for strategic choice.

To illuminate the difference between parametric and strategic choice, consider a simple illustration. Jane must choose whether to go to Ann's party. She wants to go, but only if Brian will not be there. In case one, Jane expects Brian to go to Ann's party unless his father needs him to deliver pizza, a matter having nothing to do with Jane. Whether Brian is needed to deliver pizza is, for Jane, an unknown but fixed circumstance. If she considers it likely that he is needed, she will choose to go to Ann's party; if she considers it unlikely, she will choose not to go. Jane faces a problem of parametric choice.

In case two, Brian must also choose whether to go to Ann's party. He does not want to go unless Jane will be there. Here Jane chooses on the basis of her expectation of Brian's choice, and Brian chooses on the basis of his expectation of Jane's choice. Thus Jane chooses on the basis of her expectation of a choice based on an expectation of her choice, And Brian chooses similarly. Each faces a problem of strategic choice.

Rawls relates the principles of justice, not to strategic, but to parametric choice. This may seem surprising, since he supposes that the principles would be agreed to by all rational persons in the original position. And so it may seem that each seeks to relate his choice of principles to the choices of others who are them-

[8] For this distinction between parametric and strategic, see Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge: Cambridge University Press, 1979), pp. 18–19, 117–23.

selves seeking to relate their choices to his choice. But Rawls emphasizes that this appearance of strategic interaction is misleading.[9] Behind the veil of ignorance, persons are identically situated, not only in their objective circumstances but also subjectively, in that each is completely ignorant of his capacities and interests and so is unable to distinguish himself from his fellows. They have, then, no basis for bargaining with one another, and agreement on the principles of justice may be represented by the choice of a single representative individual. The problem of rational choice to be solved is therefore one of individual decision under extraordinary uncertainty. And this is a problem of parametric choice.

I relate moral principles to strategic choice. As I shall argue, moral principles direct choice in cooperative interaction in which each person, fully aware of his or her particular circumstances, capacities, and concerns, seeks to relate his or her actions to those of others in ways beneficial to each. The rationale for moral principles — and the irrationale, we may say, for egoistic principles — emerges from an examination of the structure of such interaction.

At the end of our enquiry I shall return to this difference between Rawls and myself — a difference that also distinguishes my contractarian approach to moral theory from the utilitarian argument of John Harsanyi.[10] I suggest that there is a deep incoherence in the attempt to relate moral principles to parametric choice, since parametric choice does not fully accommodate the interaction of rational beings. Strategic rationality, which focusses on this interaction, is not fully egoistic, and moral theory is properly based on the failure of strategic egoism.

3. Once again I have indicated a destination. At the end of our brief journey we should understand more clearly both why

[9] See *A Theory of Justice,* p. 139.

[10] See John C. Harsanyi, *Essays on Ethics, Social Behavior, and Scientific Explanation* (Dordrecht and Boston: Reidel, 1976), especially ch. II and ch. VI, sections 1–5.

egoism fails and how morality relates to strategic choice. At the beginning we must relate egoism to parametric rationality.

I shall assume that rational parametric choice may be represented by a simple maximizing model. That is, I shall suppose that a parametrically rational actor behaves as if he is maximizing the expected value of some function defined over the possible outcomes of his choices. For the model adequately to represent risky and uncertain choices, in which the actor does not know the outcome of each possible choice but rather is able to assign to each only a probability distribution over possible outcomes,[11] the function must be uniquely defined up to a positive linear transformation, so that it affords an interval measure of the outcomes. A familiar example of an interval measure is temperature; the zero-point and the unit may be selected arbitrarily, but once selected the unit is constant.

I shall not ask whether rationality in parametric choice is fully captured by maximization. For our purposes we need not decide whether an actor is rational insofar as he maximizes, without consideration of what he maximizes. Thus I shall take maximization only as necessary for parametric rationality. But I shall make one further, crucial assumption — that the value maximized by an actor is relative to him or her. If Mary voted for Reagan and Harry for Carter, then we may suppose that Reagan being President had greater expected value than Carter being President as an object of Mary's choice, but lesser expected value as an object of Harry's choice.

This assumption, that value is relative and that choice is based on actor-relative value, may be related to strategic as well as to parametric rationality. Although we shall find that rational strategic choice may not always be represented by a simple maximizing model, yet strategically rational actors may be considered as assigning values to the possible outcomes. In interaction the in-

[11] We speak of risk if the probabilities are objective, of uncertainty if they are subjective.

terval measures defined by the several actors over possible outcomes are logically independent one from another. Brian's most valued outcome may be, and indeed is, Jane's least valued outcome. An outcome has no single value but a set of values, one for each actor, or indeed for each person affected by it, and there is no relationship a priori among the members of the set.

The egoist, whom we kept in a secondary role in our discussion of rational choice, has now reappeared, slightly disguised, as a species of parametrically rational actor. We first introduced the egoist as the person who pursues his own greatest happiness. He is a maximizer, albeit of a rather specific quantity — his own happiness. But for our purposes we may generalize from this characterization and think of the egoist as maximizing whatever actor-relative value he pleases — perhaps his own happiness, perhaps not. He is then simply the person whose interests, whatever they may be, have no necessary link with the interests of his fellows, so that his values provide a measure of states of affairs quite independent of their values. This generic account affords a very weak characterization of an egoist, and indeed even an excessively weak one, since it admits to the egoistic ranks persons whose interests are other-directed, provided only that their other-directed interests are not simply dependent on others' interests. But it is all that our argument will require. Our egoist then is simply a maximizer, or would-be maximizer, of actor-relative value. He satisfies the necessary condition of parametric rationality.

Before proceeding to face the egoist with the problems of interaction, I should note, out of fairness to G. E. Moore, that in introducing the maximization of actor-relative value I have already embraced what to him was the contradictory feature of egoism. For Moore, "The *good* of [something] can in no possible sense be 'private' or belong to me; any more than a thing can *exist* privately or *for* one person only."[12] Moore could allow

[12] *Principia Ethica,* p. 99.

that a state of affairs might further the well-being of one person but not that of another. He could allow that a state of affairs might be good in that it furthered one person's well-being and bad in that it hindered another's. But he denied that a state of affairs could be good in relation to the person whose well-being it furthered and bad in relation to the person whose well-being it hindered. Rather he insisted that it must be good absolutely insofar as it hindered one person's well-being and bad absolutely insofar as it hindered another's.

Moore's position might be formulated as a claim about the universality of reasons for choosing or acting. On this view, for any states of affairs P and Q, if there is a person X who is able to choose between P and Q and has a reason for choosing P over Q, then any person *Y* able to choose between P and Q has a reason for choosing P over Q. This position is embraced by many philosophers other than Moore, such as R. M. Hare and Thomas Nagel, but in embracing actor-relative value I propose to ignore it.[13]

On my view reasons for choosing have only a weaker universality. For any states of affairs P and Q, if there is a person X who is able to choose between P and Q and has a reason for choosing P over Q, then there is some relation R holding among X, P, and Q, such that, for any person *Y* who is able to choose between P and Q, (i) R need not hold among Y, P, and Q, but (ii) if R does hold among *Y*, P, and Q, then *Y* has a reason for choosing P over Q. The actor-relativity of reasons is assured by founding them on a relation between the actor and the objects of choice that does not hold for every person by virtue of holding for some person.

Suppose that Moore and I agree that enhancing my prospects of survival is a reason for me to choose to have a site for the dis-

[13] For Hare, see *Moral Thinking* (Oxford: Clarendon Press, 1981), especially chs. 5–7. For Nagel, see *The Possibility of Altruism* (Oxford: Clarendon Press, 1970), especially ch. X.

posal of nuclear wastes located in the Antarctic rather than in Allegheny County. We might expect that for Moore this would instantiate the claim: for all persons X, Y and all states of affairs P, Q, if P affords X greater survival prospects than Q, Y has a reason for choosing P over Q. For me it instantiates the claim: for all persons X and states of affairs P, Q, if P affords X greater survival prospects than Q, then X has a reason for choosing P over Q.

If Moore were right, and reasons were not actor-relative, then the maximization of some actor-relative measure of possible outcomes would be an irrational basis for choice. Not only egoism, but the entire edifice of the standard theory of rational choice, the theory that characterizes parametric rationality, would collapse. This affords an easy, but in my view unpersuasive, refutation of egoism. In granting actor-relative value, I concede egoism the initial stage of the argument concerning its rationality.

4. We are now ready to ask: what happens when the egoist, or more generally the parametrically rational actor, finds himself interacting with others of his kind? Does the endeavour to maximize actor-relative value involve him in contradiction? Or inconconsistency? Or some other form of irrationality? Is it always possible for him to put his egoism into practice? And if, or when, it is possible, is it always rational, or at least not irrational, for him to maximize?

In answering, or trying to answer, these questions, we must focus, not on interaction in general, but on strategic interaction. Were the egoist not faced with strategic problems, problems in which he seeks to adapt his choice to the choices of others adapting their choices to his, the issues we shall raise would not appear. To the extent to which interaction is not conceived in strategic terms, egoism seems fully, and perhaps even paradigmatically, rational.

This is an historically important consideration. For the most thoroughly studied form of interaction, that which occurs in the perfectly competitive market, is parametric and not strategic in character. Although each actor in the market is interacting with others of his kind, yet each chooses in a fixed environment. The firm seeks to maximize profits given known costs of factor supply and known prices reflecting aggregate consumer demand. The consumer seeks to maximize the value of his commodity bundle given known commodity prices. Since choices have fixed and indeed known outcomes, market interaction may be represented by a model that dispenses with interval measures of those outcomes in favour of weak orderings. A world conforming in every detail to the ideal of the perfectly competitive market would not raise the problems that we shall examine. Egoism is rational within the framework of the market, as Adam Smith implicitly recognized in his doctrine of the Invisible Hand, and the modern appeal of egoism is not unrelated to the dominance of the market framework in our practical thought.

But not all economic behaviour is perfectly competitive, and not all behaviour is economic. The market adequately models only a limited range of interaction. In adopting the title *Theory of Games and Economic Behavior*, Von Neumann and Morgenstern were calling attention to the insufficiently understood strategic dimension found in most interaction.[14] And it is this dimension that interests us, as we examine the problems that arise in attempting to extend the simple maximizing model of parametric rationality to accommodate strategic choice.

We shall discover two principal and distinct issues. The first is expressed in the claim that egoism is *inconsistent* — or unable always to give consistent guidance to choice. The second is expressed in the claim that egoism is *self-defeating* — that egoists fall farther short of their objectives than do some non-egoists. The first charge, we shall find, has no simple resolution. The second

[14] John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944), is the seminal work from which studies of strategic rationality have developed.

charge will be sustained, and in sustaining it we shall come to the constructive problem to which our argument is propaedeutic — to the development of moral theory as part of rational choice. We shall then understand why the strategically rational actor must be, or at least must become, a moral actor.

Let us now illustrate the two charges that we shall assess. I shall then spend the remainder of this part in examining inconsistency, leaving self-defeatingness to its successor.

The claim that egoism is inconsistent may be illustrated by our original example of strategic choice. Jane and Brian must each choose whether to go to Ann's party. Each, we suppose, has two and only two choices — to go, or not to go. If both choose to go, then Jane has chosen wrongly; she wants to go to the party, but only if Brian is not there. If neither chooses to go to the party, then Jane has also chosen wrongly; she wants to go to the party if Brian is not there. If one chooses to go and the other chooses not to go, then Brian has chosen wrongly; he wants to go to the party if and only if Jane is there. Whatever Jane and Brian choose, one of them fails to maximize his or her value. Hence one has failed to satisfy the requirements of egoism. But this failure is unavoidable. The requirements, then, can not always be satisfied. Egoism, and indeed the maximization of actor-relative value, is inconsistent.

The claim that egoism is self-defeating may be illustrated by an example long familiar among game theorists and now widely known to philosophers — the Prisoners' Dilemma. Jack and Zack are prisoners charged with a serious crime; each must choose between a confession that implicates the other and non-confession. If only one confesses, he is rewarded for turning state's evidence with a light sentence, while the other receives the maximum. If both confess, each receives a heavy sentence, but short of the maximum. If neither confesses, each will be convicted on a lesser charge and receive a sentence slightly heavier than that which would reward turning state's evidence. Jack reasons that, if Zack confesses, then he avoids the maximum sentence by confessing himself, whereas if Zack does not confess, then he gains the lightest sentence by confessing. Whatever Zack does, Jack does better to confess. Zack of course reasons in a parallel way. Given that neither is able to affect the other's choice by his own, each does better to confess, whatever the other may choose to do. Jack and Zack each maximizes his value by confessing. Each receives a heavy sentence. If neither had confessed, each would have received a lighter sentence. Jack and Zack have both satisfied the requirements of egoism and have reached a mutually costly outcome. The requirements, then, should not always be satisfied. Egoism is self-defeating.

5. To charge egoism with both inconsistency and self-defeatingness may seem excessive. If egoism fails in that it makes demands that can not be met, then why consider whether those demands are also self-defeating? The answer, of course, is that the charge of inconsistency does not affect every situation in which persons may endeavour to act egoistically. Only in some interactions, such as that of Jane and Brian, does egoism fail to direct choice.

We must be clear about the nature of this failure. Either Jane or Brian does not realize her or his most preferred outcome, but this is not sufficient to show failure of choice. If Brian chooses to go to Ann's party, then Jane, whatever she chooses, can not realize her most preferred outcome, which is to be at the party without Brian. If Jane chooses to stay home, then Brian, whatever he chooses, can not realize his most preferred outcome. In these cases, one person's most preferred outcome is excluded by the other's choice. Failure to realize one's most preferred outcome thus need not show that one has chosen wrongly, and does not in itself raise a problem for egoism. Some persons may take egoism to be inconsistent because egoists have incompatible objectives, so that not all can succeed. But the mere existence of incompatible objectives does not prevent any individual from doing as well for himself as possible, where what is possible must be determined in part by the choices of others. However, in the situation we are examining, either Jane or Brian fails to do as well as possible given the other's choice, and it is this failure, to choose what will maximize one's value given the possibilities left open by the choices of others, that is at the root of the charge that egoism is unable always to give consistent directives.

The inconsistency of egoism thus seems to arise in the following way. The egoist would be a maximizer of actor-relative value in strategic interaction. What is required for one to be such a maximizer? It would seem that one must always choose what maximizes one's value given the choices of the others. Our example shows that it may be impossible for everyone to make such a choice. *Any* person can make such a choice; given the choices of others any person has a maximizing alternative. But not *every* person can make such a choice. Not everyone can always be a maximizer of actor-relative value in strategic interaction. Egoism, in requiring this, is inconsistent.

This argument moves too quickly. In a world of risk and uncertainty, even the parametrically rational actor can not ensure that he maximizes actor-relative value. Given his estimate of the probability of alternative circumstances, he can maximize *expected* actor-relative value, but in choice he can set his sights no higher. Similarly, the strategically rational actor must be satisfied if he maximizes expected value. And to do this, he need not always choose what maximizes his value given the choices of the others, but only what maximizes his value given the choices he expects the others to make. If Jane supposes it unlikely that Brian will choose to go to the party, then she may maximize her expected value by choosing to go. If Brian supposes it likely that Jane will choose to go to the party, then he may maximize his expected value by choosing to go. Jane will then be disappointed by the outcome, but her choice, it may seem, satisfies the requirements of egoism. We have found no reason to claim that not everyone can be a maximizer of expected actor-relative value in strategic choice, even if some must be disappointed by the outcome.

But this rejoinder also moves too quickly. Let us suppose that Jane and Brian know each other to be would-be maximizers of actor-relative value. Then for each to maximize expected value, each must choose on an expectation about the choice the other will make based on an expectation about what his or her own choice will be. If Jane chooses to go to the party, she does so expecting Brian to choose not to go because he expects her to choose not to go. If Jane chooses to stay home, she does so expecting Brian to choose to go because he expects her to choose to go. Whatever she chooses, Jane must base her choice, if it maximizes her expected value, on an expectation that requires Brian to have a mistaken expectation about her choice. And similarly, Brian must base his choice on an expectation that requires Jane to have a mistaken expectation about his choice.

We may now give a more satisfactory explanation of the failure that seems to make egoism inconsistent. The egoist would maximize expected actor-relative value in strategic choice. Thus he must seek to maximize his value given the choices he expects to be made by others who seek to maximize their values given the choices they expect to be made by others, himself included. But the following three propositions can not all be true:

1. An egoist always chooses to maximize value given the choices he expects others to make.
2. An egoist always expects other egoists to choose to maximize value given the choices they expect others, himself included, to make.
3. In satisfying 1 and 2, an egoist is never required to suppose that the expectations of other egoists are mistaken.

The failure of egoism thus lies in the necessity of attributing mistaken expectations to others in situations such as that of Jane and Brian, in order to suppose that each person chooses to maximize actor-relative value given the choices he expects the others to make.

Let us introduce some useful terminology for expressing what we have argued. An action maximizing the actor's value in interaction with others is a *best response* to the others' actions. An egoist chooses an expected best response. In some situations no set of actions, one for each person, is a set each member of which is a best response to the other members. In the terminology of the theory of games, a set of mutual best responses is a *Nash-equilibrium* set;[15] in some situations there is no Nash-equilibrium set of actions. In such situations egoists can all choose expected best responses only if some have mistaken expectations. The existence of a Nash-equilibrium set of actions is a necessary condition for successful and informed egoistic choice.

Let us treat a *principle for choice* as a function that takes sets of alternative actions into subsets of themselves. (For any set S, the corresponding subset is then termed the *choice set, C(S).)* A principle is *complete* for any domain if and only if it takes each member of the domain into a non-empty subset. A principle is *egoistic* only if it takes each set S into a subset *C(S)* , the members of which maximize some measure defined over S. In a community of sufficiently informed egoists, a principle that determines a choice for each person involved in an interaction must determine choices that maximize each person's value given the other choices it determines. In other words, a principle that includes in its domain all of the sets of alternative actions making up an interaction must take each set into a subset which has as members only actions belonging also to Nash-equilibrium sets for the interaction. Since for some interactions there is no Nash-equilibrium set, there can be no egoistic principle for choice defined over the domain consisting of all sets of alternative actions in all possible interactions. There can be no egoistic principle of choice complete for all strategic interaction. This gives precise sense to the accusation that egoism is inconsistent.

[15] The term "Nash-equilibrium" refers to John F. Nash, who is responsible for the core result concerning equilibrium in strategic interaction to be discussed in the next section.

6. Having called the resources of the theory of rational choice to our aid, we now find that they open unexpected complexities in our attempt to assess the consistency of egoism. Only the first round of our discussion is completed; we begin the second round by turning from *actions* to *strategies.* A strategy is a lottery or probability distribution over possible actions. To this point we have thought of each actor choosing among possible actions; let us now enlarge the choice space and think of each actor choosing among possible strategies. To choose an action is in effect to choose a strategy assigning that action a probability 1 and each alternative a probability 0. Such a strategy is termed *pure.* But there are countless *mixed* strategies which assign a positive probability to each of two or more alternative actions.

We have supposed that each actor may be represented as seeking to maximize the value or expected value of a function that provides an interval measure of possible outcomes. The value assigned to each *action* is the weighted sum of the values of its possible outcomes, where each weight represents the probability of the outcome given performance of the action. The value assigned to each *strategy* is then the weighted sum of the values of its possible actions, where each weight represents the probability assigned to the action by the particular strategy. We now suppose ‘that the egoist seeks to maximize some actor-relative value in choosing among his possible strategies.

With this supposition we may, surprising as it might seem, rescue the egoist from the charge of embracing an inconsistent basis of choice. For more than thirty years ago John F. Nash proved that, in any interaction among finitely many persons, each with only finitely many actions or pure strategies, there is at least one Nash-equilibrium set of strategies.[16] Or in other words, there is at least one set of strategies, one for each actor, each of which is a best response to the other members of the set. And the exis-

[16] See John F. Nash, "Noncooperative Games," *Annals* of *Mathematics 54 (1951),*pp. *286-95.*

tence of such a Nash-equilibrium set satisfies our requirement for egoistic choice that is both maximizing and correctly informed.

If we would apply the existence of a Nash-equilibrium set of strategies to resolve the problem of choice facing Jane and Brian, we must provide each with an interval measure of possible outcomes. Rather than doing this and solving the resulting mathematical problem, we shall develop intuitively the idea of determining a pair of strategies each of which is a best response to the other. Suppose that Jane despairs of concealing her strategy choice from Brian. She expects that, should she select a strategy giving a high probability to going to Ann's party, Brian will respond by choosing to go to the party so that the likely outcome will be undesirable for her. And she expects that, should she select a strategy giving a low probability to going to Ann's party, Brian will respond by choosing not to go to the party, so that once again the likely outcome will be undesirable. What then is she to do? She needs a strategy that leaves Brian indifferent between choosing to go to the party and choosing not to go, that affords him the same expected value whatever he chooses. Similarly, Brian needs a strategy that leaves Jane indifferent between choosing to go to the party and choosing not to go. If Brian is indifferent as to his choice of strategy, then any strategy is a best response for him; similarly, if Jane is indifferent as to her choice of strategy, then any strategy is a best response for her. Therefore if each chooses a strategy that leaves the other indifferent, each strategy must be a best response to the other, so that the pair constitutes a Nash-equilibrium set.

Jane does not first form an expectation about Brian's choice of strategy and then choose her best response to it. Instead she chooses a strategy that leaves Brian nothing to choose among his responses. And Brian chooses a strategy that leaves Jane nothing to choose among her responses. In situations such as the one we are considering, it is always possible to find a strategy that leaves the other indifferent, and such strategies are mutual best responses, so that successful egoistic choice seems possible.

If we make not implausible assumptions about the relative values, to Jane and to Brian, of the possible outcomes of their choices, we might find that Jane should choose a mixed strategy with probability 217 of going to the party and 5/7 of not going, and that Brian should choose a mixed strategy with probability 3/5 of going to the party and 2/5 of not going.” And these strategies would constitute the unique Nash-equilibrium pair. If either were able to calculate one of these strategies, she or he would have sufficient knowledge of the situation to calculate the other. Neither Jane nor Brian need be concerned to conceal the choice of strategy from the other. Of course, at some point each must determine what actually to *do* — no doubt using a handy pocket randomizing device appropriately programmable for any lottery. We must suppose that the outcome of this determination remains unknown to the other until the action is actually carried out. In supposing that each chooses a strategy, we suppose that each considers the other's choice of a strategy, forming expectations about it but not any more determinate expectations. If Brian could know that Jane's handy randomizer said “Go!”,then, rather than consulting his own, he would simply head for Ann's party.

We have no reason to assume that Jane and Brian actually have the information about each other's values needed to calculate strategies in Nash-equilibrium, And this is a very simple interaction, In more complex situations the procedure required to determine strategies in Nash-equilibrium may be more difficult,

[17] These mixed strategies yield equilibrium for the following case. Arbitrarily assigning the value 1 to an actor's most favoured outcome, and *0* to the least favoured outcome, we find that an interval measure of Jane's preferences assigns 1 to going to Ann's party if Brian does not, 1/2 to not going if Brian goes, 1/4 to not going if Brian does not, and *0* to going if Brian goes. And we find that an interval measure of Brian's preferences assigns 1 to going to Ann's party if Jane goes, 1/2 to not going if Jane does not, 1/10 to going if Jane does not, and 0 to not going if Jane does. Jane's mixed strategy affords Brian an expected utility of 5/14 whatever he does, and Brian's mixed strategy affords Jane an expected utility of 2/5 whatever she does. Note that the utility values for Jane and Brian are *not* interpersonally comparable; we may not infer that Jane may expect to do better from the situation than Brian from the fact that 2/5 is greater than 5/14.

even if the information needed is available. Me know from Nash's proof that there must be at least one Nash-equilibrium strategy set, but this knowledge may have no practical application. Thus I make no claim about the ability of actual egoists to choose best response strategies. But there is a fundamental difference between recognizing that failure does occur and demonstrating that it *must* occur. There is a principle for choice among strategies that includes in its domain all of the sets of alternative strategies making up an interaction, and that takes, as values, sub-sets each of which has as members only strategies belonging also to Nash-equilibrium sets for the interaction. Egoists are no longer set a task that is insoluble in principle.

To avoid possible misunderstanding, note that the strategic consistency of egoism can in no way affect the impossibility, in some situations, of actually selecting only actions that meet the egoistic requirement. When Jane and Brian actually act, and discover what each other does, then one will not maximize value given the other's behaviour. Moving to the strategic level does not enlarge the actual possibilities for action, and so does not affect the impossibility of successful informed maximization by both Jane and Brian in terms of their actions. But if each selects a strategy that is a best response to the other's selection, then each will know that whatever the outcome, she or he maximized expected value. Neither will judge his or her choice to have failed *as a choice.*

Before we conclude this round of our argument, we should admit that we have not shown the existence of a principle for egoistic choice among strategies that includes all sets of alternative strategies in all interactions in its domain. All that we have shown is that the requirement that the strategies selected by the principle for any interaction form a Nash-equilibrium set can be satisfied. But we must not suppose that a sufficient principle of egoistic choice would simply require each actor to select a strategy — any strategy — belonging to such a set. For although this would suffice in the simple situation,we have considered, in which there is but one Nash-equilibrium pair of strategies, yet in other more complex situations there may be a multitude of sets, each of which contains only strategies that are best responses to each other, but such that a strategy belonging to one Nash-equilibrium set is not a best response to strategies belonging to other Nash-equilibrium sets. Consider, for example, a situation in which several persons want to meet but are indifferent among several possible meeting-places. If each chooses to go to the same possible meeting-place, then each action is a best response to the others; if each chooses to go to a different meeting-place, then the actions are not best responses. We have not considered how egoists, embarrassed by such riches, would select among different sets of actions or strategies in Nash-equilibrium. Thus the present round of our argument concludes only with the judgement that the accusation of inconsistency against egoism is not proven. We have a Scots verdict.

7. An exhaustive examination of the problems created for egoists by the existence in some situations of several sets of actions or strategies, each in Nash-equilibrium, is beyond the scope of our present enquiry. I shall focus on but one such problem, arising from the plausible requirement that egoists coordinate their choices to bring about a mutually superior Nash-equilibrium, should one exist and should each stand to lose from the failure to coordinate.[18]

Let us begin by considering a simple game. Two players are each given a coin and must choose whether to show heads or to show tails. No communication between them is permitted; each must choose in ignorance of the other's choice. If both show the same, then each wins a sum of money, but the sum is larger if

[18] The problem discussed in this section is essentially the same as that discussed in my paper "The Impossibility of Rational Egoism," *Journal* of *Philosophy* 71 (1974), pp. 439-56. This earlier paper examines certain details not treated here, but focusses less clearly on the issue identified here as the consistent application of a principle for choice to an interaction and its sub-interactions.

both show heads. If one shows heads and the other shows tails, then each loses a sum of money. In this game there are two pairs of actions in Nash-equilibrium — each showing heads, and each showing tails. But the former pair is a superior equilibrium, dominating the latter, since the outcome if each shows heads has greater value for each player than the outcome if each shows tails. A principle of choice for egoists must surely accommodate this. We might initially suggest that such a principle must require the selection of strategies that will ensure coordination on a superior equilibrium, should there be one and should it satisfy certain accessibility considerations that we may ignore here.[19] Thus we suppose that in this game rational egoists choose heads.

A variant on our game may suggest that the proposed coordination requirement is too strong for egoists. Suppose that each player gains a sum of money if the other shows heads and loses an equivalent sum if the other shows tails. In this game every pair of strategies is in Nash-equilibrium, since each player is entirely indifferent about his own choice; what he gets is determined by what the other does. There is a unique equilibrium superior to all others, arising if each player shows heads. But the requirement that players coordinate on this equilibrium is egoistically unmotivated. Neither player has any incentive to show heads, since showing tails would neither reduce his expected value nor affect the occurrence of equilibrium. In this game we have no reason to suppose that rational egoists would choose heads rather than tails.

Intuitively, we want to treat coordination on strategies belonging to a set in superior Nash-equilibrium as an egoistic require-

[19] Consider a situation with three outcomes resulting from sets of strategies in Nash-equilibrium. Let the outcomes be P, Q, and R, and let P and Q be indifferent (from the standpoint of each individual) but superior to R. If communication is impossible, and if neither P nor Q possesses any naturally salient feature, then coordination may be possible only on the inferior equilibrium R, because its very inferiority distinguishes it, whereas nothing distinguishes P from Q. Here P and Q are effectively inaccessible; without communication neither can be singled out as a target for coordination.

ment only if defection from such coordination would reduce the defector's expected value. I shall not however attempt to formulate this requirement precisely, since it will be clear, in the situation we are about to consider and that poses a problem for the consistency of egoism, that coordination is egoistically motivated.

Consider now a more complex game, in which three players, A, B, and C, are each given a coin, and must choose, without communication, whether to show heads or to show tails. The values, or payoffs, of the possible outcomes of the different combinations of actions are shown in this table:

Action (= Pure strategy)			*Payoff*		
A	*B*	*C*	*A*	*B*	C
H	H	H	$1.00	$1.00	$1.00
H	H	T	$1.50	$1.50	*0*
H	T	H	-$1.50	-$1.50	*0*
T	H	H	-$1.50	-$1.50	*0*
H	T	T	$1.50	*-$3.00*	$1.00
T	H	T	*-$3.00*	$1.50	$1.00
T	T	H	$1.50	$1.50	*0*
T	T	T	$1.00	$1.00	$1.00

In this game there is a single set of strategies in Nash-equilibrium; each player shows heads, This is easily verified; showing heads is each player's best response if the others show heads, so the set is in equilibrium. It is unique since, given any other set, at least one player would do better to change her response, so the other set of strategies is not in Nash-equilibrium.

Let us suppose then that A expects C to show heads. For A reasons that if expectations are correct, and if each choice is a best response to the others, then the strategies must be in Nash-equilibrium, and showing heads yields the unique Nash-equilibrium. But then she notes that *if* C shows heads, then she and B would each do better were they to show tails. For then they would take C's winnings and add them to their own, gaining $1.50 instead of $1.00. Taking C's choice as fixed by the requirement of equilibrium, and focussing then solely on the interaction between A and B, the payoffs for the possible outcomes are shown in this table:

Action		*Payoff*	
A	*B*	*A*	*B*
H	H	$1.00	$1.00
H	T	-$1.50	-$1.50
T	H	-$1.50	-$1.50
T	T	$1.50	$1.50

In this sub-game there are two pairs of strategies in Nash-equilibrium — each player shows heads, and each shows tails. But the latter pair dominates the first; it is a superior equilibrium. And coordination on it is egoistically motivated; each stands to gain from achieving coordination and to lose if she defects from it. If A expects B's reasoning to parallel her own, then she concludes that, given that C may be expected to show heads, then she should show tails with the expectation that B also will show tails.

But C's deliberation need not have ceased with the realization that the requirement of equilibrium determines that she show heads. For if she correctly anticipates the reasoning of A and B, leading them to coordinate on tails, then she must conclude that she too should show tails. If she expects them to show tails, then showing tails is her best response, enabling her to keep a gain of $1.00 rather than losing it to A and B. But then if A and B anticipate this further deliberation by C, they should coordinate on heads; expecting C to show tails, they realize that their strategy pair showing heads now dominates the pair showing tails, since it enables them to recapture C’s gain. And if C anticipates this further deliberation on the part of A and B, then, expecting them to show heads, she too should show heads. Thus she returns to the set of strategies in Nash-equilibrium, the point of departure for the circle that we have traced.

Crucial to the argument implicit in our discussion of this game is a claim about the consistency required for a principle of choice to be successfully employed. Suppose that a principle includes in its domain all of the sets of strategies constituting an interaction. Thus for each actor it yields a sub-set of his strategies as his choice set. Let each sub-set contain a single strategy; this will arise if the principle satisfies the Nash-equilibrium requirement and the situation has a unique Nash-equilibrium. Suppose that one actor chooses the unique strategy in his choice set, as the principle requires. Taking that choice as a fixed circumstance, apply the principle to the reduced interaction among the remaining actors. Then our claim is that, if the principle is consistent, it must yield, for each remaining actor, a choice set that contains the strategy in his original choice set. The principle must yield consistent guidance, whether an actor apply it directly to his choice of strategy in an overall interaction, or whether, taking for granted that some others will conform to it, he apply it to his choice of strategy in the resulting sub-interaction. A principle that says, “Everyone should show heads, but if actor C shows heads then everyone else should show tails,” is inconsistent.

If we accept this view of consistency, then no egoistic principle of action can be both complete and consistent. An egoistic principle must satisfy the equilibrium requirement, that strategies chosen in an interaction be mutual best responses, and the coordination requirement, that strategies chosen yield a superior equilibrium, if one exists and defection from it would be costly to the defector. A complete principle of choice for interaction must yield a non-empty choice set for each set of strategies in each possible interaction. A consistent principle of choice must yield compatible choice sets when applied to an interaction as a whole and to any reduced sub-interaction resulting from taking its application to some of those interacting as given. Our game of matching coins shows that egoism, completeness, and consistency are jointly incompatible.

Egoists seek individually advantageous responses to the actions of their fellows and mutually beneficial coordination among their actions. These goals prove to be in conflict in situations such as the game we have discussed. Of course, the failure to attain a goal can be accepted. But we have shown that in some situations, some actors can not do as well for themselves as possible, given what the others do. If A, B, and C do not all show heads, then at least one could do better given the possibilities left open by the others' choices. If A, B, and C do all show heads, then A and B could coordinate on a mutually better outcome. In such a situation, egoism makes inconsistent demands, and so fails.

8. So what can an egoist do? Why, he can do his best. But we now know that this answer misleads. We have discovered situations in which not everyone *can* do his best. *Any* person can do his best, but *not every* person can. If we ask, what can *egoists* do, we must not reply that *they* can do their best.

Let us change the question. What can egoists choose? If we take strategies as the objects of choice, then we can answer: they can choose their best. But again we know that the answer misleads. We have discovered situations in which everyone can choose her best, given all other choices, but only if some fail to coordinate their choices on what is, for them, mutually best. And were they to succeed in coordinating, then some individual would fail to choose her best. Taken individually egoists can choose their best; taken, let us say, coordinatively, not all egoists can.

Faced with the complexities of strategic interaction, the egoist must soon lose the naïve hope of formulating a complete and consistent principle for choice satisfying the conditions implicit in his egoistic stance. The problem proves less tractable than either philosophical critics or proponents of egoism have recognized. Perhaps, then, the first lesson for the would-be egoist is to place less trust in the words of philosophers and pay closer attention to the structures of interaction exhibited by game theorists. But the message our game-theoretic enquiry conveys must surely dishearten him: anyone may do his best, but not everyone.

Yet may not his dismay and puzzlement remain only that? The demonstrably impossible is, simply, impossible. The egoist can do his best. That the structures of interaction constrain doing one's best in initially unexpected ways neither contracts nor expands the real horizons that egoists, in their actions and choices, have always faced.

This would dismiss too easily the import of our argument. Egoists, and not egoists alone but all would-be maximizers of actor-relative value, have been on the whole unaware of the structure of their predicament. They have recognized problems arising from the incompatibility of their professed values; some have interpreted this incompatibility as a sign of irrationality, some have seen it only as the basis of inevitable frustration. They have not, however, recognized the constraints that exist on doing, and choosing, one's best. They have not recognized that the very correctness of the expectations persons may form about the choices of those with whom they interact may ensure that someone must fail to *do* what would maximize his value, given the possibilities left open to him by those expected choices. They have not recognized that a full awareness of the possibilities for advantageous coordination may ensure that either someone fails to choose what would maximize his value given the choices of others, or that some fail to choose what would be mutually maximizing given the choices of others.

A principle that prescribes a choice for each and then, on the assumption that some follow it, prescribes a different choice for the others, forces more than dismay and puzzlement on those who would adhere to it, The inconsistency of egoistic principles requires us to think again about certain failures of interaction, to reappraise what goes wrong in the light of the inescapable nature of certain conflicts. Even if anyone can do his best, the fact that not everyone can do his best forces us to reconsider the attribution of responsibility for failure in situations comparable to those we have examined. We excuse, or partially excuse, a person's failure to achieve his objective, if we find that he did his best; must we now excuse a person's failure to do his best, if we find that not everyone could?

Here I leave these and other implications of our discussion of the inconsistency of egoism to the reader's reflection. Perhaps, just as we found a partial remedy for the failure of egoistic choice among actions by considering choice among strategies, so we might be able to find a partial remedy for the conflict between individual maximization and mutual coordination. We must not draw too firm a conclusion to our treatment of the consistency of egoism, and so we must hesitate in assessing its implications for such issues as the attribution of responsibility. But we put these issues aside in part because an even more pressing question awaits us. If there are limits to what egoists can do and can choose, yet in many situations all can indeed do their best. But should they? The would-be egoist who as yet sees no reason to change his ways may yet have to reconsider if those ways can be shown to be self-defeating.

II. WHAT SHOULD AN EGOIST DO?

1. "The very *raison d'être* of a morality is to yield reasons which overrule the reasons of self-interest in those cases when everyone's following self-interest would be harmful to everyone,''[20] As a claim about actual moralities, this statement by Kurt

[20] Kurt Baier, *The Moral Point of View: A Rational Basis of Ethics* (Ithaca, N.Y.: Cornell University Press, *1958),* p. 309.

Baier may well be false, or at most a very partial truth. But as a claim about rational morality — about a morality that would be acceptable to rational actors — this statement is, I believe, the exact truth. A rational morality is a constraint, or set of constraints, on the maximization of actor-relative value with which it is rational for would-be maximizers of such value to agree and comply.

But how can it be rational for maximizers to constrain their maximizing activity — or, more specifically in terms of our enquiry, for egoists to constrain their egoism? I propose to answer this question. But for some years I thought no answer was possible. Indeed, I said as much.

When I first considered Kurt Baier's conception of morality, I found myself trying to understand the conflict between reasons of self-interest and overriding reasons, and I wrote, and read, a paper in which the issues became obscured in a labyrinth of words. After listening to those words, Howard Sobel took me aside and, quickly sketching a matrix on a sheet of paper, said, "Look! You're talking about the Prisoners' Dilemma." And I looked, and it was as if scales fell from my eyes and I received sight."

But at first I saw poorly. I saw in the Prisoners' Dilemma a clear representation of the conflict between interested reasons and moral or cooperative reasons, but neither seemed overriding. I saw a conflict between two conceptions of rationality — the one individual and prudential, the other collective and moral. And I said that "the individual who needs a reason for being moral which is not itself a moral reason cannot have it. . . . For it is more than apparently paradoxical to suppose that considerations of advantage could ever of themselves justify accepting a real disadvantage."[22] I was wrong. It is that supposedly genuine

[21] The incident described here occurred at the University of California, Los Angeles, probably in November 1965.

[22] David Gauthier, "Morality and Advantage," *Philosophical Review* vol. 76 (1967), p. 470.

paradox that I want now to confute — to show that one can and does have a non-moral reason for being moral, a reason that must be recognized even by the egoist.

Egoism is self-defeating. The objective of the egoist is to do as well for himself as possible, to maximize actor-relative value. More typically we think of the egoist as identifying his interest or his advantage with what he values, so that his objective is simply to maximize that interest or advantage. But the egoist falls short of this objective — and falls farther short than some who are not egoists. Reflecting on his maximizing objective, the egoist finds reason to change his ways, casting *off* the egoistic scales from his eyes and seeing as a moral being — even as a being who accepts real disadvantages. The egoist, embarked on the journey of rational choice, finds, contrary to all expectation, that his destination is moral theory.

To show that egoism is self-defeating is no simple matter. As we shall see, it is not enough to show that egoists, in maximizing actor-relative value, fail to do as well for themselves *collectively* as they might. It is not enough to show, in Baier's words, that "everyone's following self-interest would be harmful to everyone." We must rather show that each person's following self-interest is harmful to himself, that each fails to do as well for himself *individually* as he might. Only an argument addressed to the individual egoist can hope to show that *his* ways are self-defeating. But we may begin from the perspective of everyone, from the failure of egoists to do as well for themselves as possible, and then show how this perspective may be linked to that of the individual. And so we may begin with the Prisoners' Dilemma,

2. If philosophers have paid little attention to considerations of Nash-equilibrium in examining the consistency of egoism, they have become quite familiar with the Prisoners' Dilemma as exhibiting the seemingly self-defeating character of egoistic behaviour. Let us review exactly what the Dilemma shows. Each prisoner — Jack and Zack as I called them in the preceding

part — has a strategy that is a best response to whatever strategy the other chooses. This strategy is confession. But the outcome if each chooses his best response is disadvantageous to both. Both would do better if both chose the alternative strategy — nonconfession or silence. Each does best to confess whatever the other does, but each does better if neither confesses than if both confess.

We need another piece of terminology to talk about the Dilemma — Pareto-optimality.[23] An outcome is Pareto-optimal if and only if no feasible alternative affords some person greater value and no person lesser value. Or, assuming a link between value and preference, if and only if no alternative would be preferred by some and dispreferred by none. An equivalent formulation is that an outcome is Pareto-optimal if and only if every feasible alternative that affords some person greater value also affords some other person lesser value.

Consider the outcomes possible for Jack and Zack. If both confess, each receives a heavy sentence, but short of the maximum. If neither confesses, each receives a light sentence, but exceeding the minimum. If one confesses and the other does not, the one confessing receives the minimum sentence and the other receives the maximum. Let us assume that their values are related inversely to the length of their sentences. Then if we consider in turn each pair of outcomes, we find that in every case Jack prefers one member of the pair and Zack the other, except that both prefer the outcome if neither confesses to the outcome if both confess. The outcome of mutual confession is therefore *not* Pareto-optimal; there is an alternative affording both greater value. Every other outcome is Pareto-optimal; for each such outcome, every alternative affording one prisoner greater value affords the other lesser value.

[23] The term "Pareto-optimality" refers to Vilfredo Pareto, who did not talk about optimality at all, but rather ophelimity.

But the strategies leading to confession are in Nash-equilibrium; each is the best response to the other. And since each is the unique best response whatever the other prisoner chooses, no other set of strategies is in Nash-equilibrium. The outcomes thus divide into two exclusive and exhaustive sets-the set of Pareto-optimal outcomes and the set of outcomes resulting from strategies in Nash-equilibrium. That the sets are exhaustive is not a common characteristic of structures of interaction. But that the sets are exclusive is common, or at least not uncommon, and represents that feature of the Prisoners' Dilemma that makes it a supposed dilemma. For if, as I have argued, informed egoists are restricted to outcomes resulting from strategies in Nash-equilibrium, then in such situations as the Dilemma, egoists are barred from Pareto-optimality. Each may succeed in doing as well for himself as he can, but everyone could do better.

An outcome may be conceived in two quite different ways, each important to rationality. On the one hand, an outcome may be conceived as the product of the members of a set of strategies; on the other hand, it may be conceived as a set of payoffs. Conceived as the product of a set of strategies, we say that it is in Nash-equilibrium if and only if each strategy maximizes the actor's value given the other strategies. Considered as a set of payoffs, we say that it is Pareto-optimal if and only if each payoff maximizes the recipient's value given the other payoffs.[24] No complete principle for choice in interaction takes every set of alternative strategies into a (non-empty) choice set, some member of which belongs also to a Nash-equilibrium set that yields a Pareto-optimal outcome. No complete principle can ensure both Nash-equilibrium and Pareto-optimality in every interaction. This is the impossibility theorem, illustrated by the Prisoners' Dilemma, that egoists and all maximizers of actor-relative value must face.

[24] This is true only if payoff functions are continuous. More generally, an outcome is Pareto-optimal if and only if each payoff maximizes the recipient's value on condition that no other payoff is decreased.

Note that this impossibility does not reveal a further inconsistency in egoism. The problem here is not similar to those discussed in the preceding part. There is no difficulty in formulating an egoistic principle for choice in the Prisoner's Dilemma and similar situations of incompatibility between equilibrium and optimality. The equilibrium requirement for an egoistic principle is straightforwardly satisfied, with no need to resort to strategies rather than to actions as the objects of choice. And the equilibrium requirement suffices to determine an egoistic principle; coordination is irrelevant in the Dilemma. The coordination requirement for egoistic principles of choice that I introduced in the preceding part applies to certain situations with more than one Nash-equilibrium strategy set. But in the Dilemma there is only one equilibrium set. Egoists, as maximizers of actor-relative value, are able to coordinate their strategies only within the limits allowed by the requirement that each actor consider his strategy to be a best response to the strategies he expects the others to choose. If, as in the Dilemma, each actor has a unique best response whatever he expects the others to choose, then coordination has no place.

Let us illustrate the difference between a simple coordination problem and the Dilemma by contrasting two games. First, consider again the two-person game of matching coins that served in Part I, section 7, to motivate the coordination requirement, here with determinate monetary values.

Action		Payoff	
A	*B*	*A*	*B*
H	H	$2	$2
H	T	–$2	–$2
T	H	–$2	–$2
T	T	$1	$1

Here each does best to show what the other shows; this assures equilibrium. Both do better if both show heads than if both show tails; mutually beneficial coordination thus enables the players to select among the equilibrium strategy sets, But now consider this Dilemma-type game:

Action		*Payoff*	
A	*B*	*A*	*B*
H	H	$2	$2
H	T	-$3	$3
T	H	$3	-$3
T	T	$1	$1

Here again both do better if both show heads than if both show tails. But this consideration never enters into an egoistic principle for choice. For each does best to show tails whatever the other does; this alone assures equilibrium. And so the equilibrium requirement leaves no room for other considerations. Egoists should *not* show heads, because showing heads does not lead either player to do as well for himself as he can. The outcome of egoistic behaviour may well be regarded by the players as unfortunate, but in choosing tails, each does his best for himself. But should he? Should he *be* an egoist?

Before attempting to answer this question, we must generalize from the particular structure of the Dilemma to the underlying conflict between Nash-equilibrium and Pareto-optimality that it illustrates. And we must not misunderstand the nature of this conflict. An egoist concerned to maximize actor-relative value is utterly indifferent to both equilibrium and optimality. He cares only for his own payoff. If the strategies of others are given, then he chooses that strategy most profitable to himself; if all behave in this way, then Nash-equilibrium is the unintended result. If the payoffs of others were given and he were to choose among payoffs, then he would choose that most profitable to himself; if all behaved in this way, then Pareto-optimality would be the unintended result. The egoist is concerned with payoffs, but since choice determines actions or strategies, he can express his concern with payoffs only in his choice among strategies. What the Dilemma reveals is that in some situations, his choice does not give effective expression to his concern. Selecting among strategies, the egoist may be unable to maximize his payoff given the payoffs of others, and so may be unable to obtain some benefit that he could enjoy at no cost to others. Let us then say that the egoist faces *strategy-payoff conflict.* This is the general problem that the Dilemma reveals.

3. How important is strategy–payoff conflict? Grant that its occurrence must complicate life for egoists and indeed for all maximizers of actor-relative value. But does it occur, except in the structures of interaction studied by game theorists? If it is a phenomenon of no practical significance, then it can hardly serve as the basis for an argument that egoism is self-defeating.

It does occur. Indeed, strategy–payoff conflict is a fundamental phenomenon of social life. It constitutes the core of the problem of ensuring the optimal supply of public or collective goods. It explains the sub-optimality that characteristically results from failures to internalize effects — the coincidence of net social costs from pollution with net individual benefits to polluters. It is at the heart of Garrett Hardin's tragedy of the commons,[25] and helps explain John Kenneth Galbraith's observation that an affluent society enjoys "private opulence and public squalor." [26] It enables us to understand why even a government that spent its funds wisely would need an I.R.S. to collect them. Free-riders and parasites flourish in the context of strategy–payoff conflict.

[25] See "The Tragedy of the Commons," *Science* 162 (1968), pp. *1243-48.*

[26] See *The Affluent Society* (Boston: Houghton Mifflin, 1958), p. 257. Galbraith does not, however, focus on this explanation.

Its importance, widely recognized today, was long obscured in much of our social and economic thought. There are two principal reasons for this. First, economists since Adam Smith have tended to focus unduly on the perfectly competitive market — from which strategy–payoff conflict is blissfully absent. As Russell Hardin has dramatically expressed it, the Prisoners' Dilemma is the back of the Invisible Hand.[27] In the perfect market the Invisible Hand ensures that if each pursues his own interest, the social interest is furthered, albeit unintentionally. We may make this more precise by saying that market activity — in which each individual seeks to maximize the value of a function defined over the goods he consumes and the factor services he provides — leads to an outcome on society's utility-possibility frontier, so that no person's position could be improved without worsening that of some other person. The equilibrium resulting after all voluntary exchanges is Pareto-optimal.

Were the world to be, as some economists of the Chicago school are alleged to suppose that it is, a perfectly competitive market, then egoists would have no reason to change their straightforwardly maximizing ways. The Prisoners' Dilemma would be a logical curiosity, revealing the possibility of interactions, happily never realized, in which egoists would fail to end up on the utility-possibility frontier and so would fail, collectively, to do as well for themselves as possible. But illuminating as the market is in showing us the possibility of interactions that give rise to no problems for maximizers of actor-relative value — indeed, illuminating as the market is in revealing to us a type of interaction that would not need to be guided by those principles, constraining maximizing behaviour, that constitute a rational morality — yet to most of us the real world does not seem to be a very close approximation to the realm of perfect competition. And so we expect to face strategy–payoff conflicts,

[27] See *Collective Action* (Baltimore: Johns Hopkins University Press, 1982), page 7.

both in our everyday interactions and in the design of the social institutions that frame those interactions.

But even when we turn away from the perfect market, we encounter a second factor that has obscured our awareness of this conflict. For awareness of the failure of the market as a model for much of our social interaction does not entail awareness of the core problem facing non-market public or collective behaviour. There is a strong temptation to suppose that, just as a rational individual will, within the limits of available information, so choose that he does as well for himself as possible, so a group of rational individuals will also choose that they do as well for themselves as possible. W e extrapolate from individual action to group or collective action.

Mancur Olson, Jr., in his book *The Logic of Collective Action*, written some twenty years ago, seems to have been the first to recognize the general fallacy involved in this extrapolation.28 Here I shall illustrate it with an example adapted, not from his work, but from that of Russell Hardin.[29] Suppose that 10 units of a pure public good in full joint supply are available to a community of 10 persons. Each unit costs $5 and affords each member of the community a benefit of $1. Each must decide whether to contribute $5 to the social provision of the good. If all contribute, total benefit is $100 and cost $50, for a net social benefit of $50 and a net benefit to each individual of $5. If no one contributes, then net social and individual benefit are both $0. Nevertheless, no one who seeks to maximize his payoff will contribute. Each reasons that *n* other persons will contribute, where *n* takes a value from *0* to 9. If he also contributes, net social benefit is $10(n + 1), divided so that net benefit to each contributing individual is $(n – *4*) and to each non-contributor $(n + 1). If he does not contribute, net social benefit is $10n divided so that net

[28] *The Logic of Collective Action* (Cambridge: Harvard University Press, 1965); the fallacy is spelled out in the Introduction, pp. 1–2.

[29] *Collective Action,* pp, 25–27.

benefit to each contributing individual is $(n – 5) and to each non-contributor $n. Thus, if he contributes, his net benefit is $(n – *4*); if he does not contribute, his net benefit is $n. For all possible values of n, $n is greater than $(n – *4*), so he chooses not to contribute.

The parallel with reasoning in the Prisoners' Dilemma should be evident. Given a pure public good, each individual chooses to ride free; of course the result is that there is nothing on which to ride. I have considered only an artificially simple case; in more realistic cases in which the value of each unit of the good diminishes as more units are obtained, it may be that some units will be bought by individuals who find it worth their while to pay the entire cost of supplying the unit to everyone, but before optimal supply is reached, each will prefer to ride free at the current level of supply rather than to contribute an additional unit.

Recognition of the problem of collective action should dispel any temptation to suppose that my argument is addressed not to us, but only to very different persons — to egoists. We are, all of us, maximizers of actor-relative value — or a near approximation thereto — in many of our interactions. And we all face the problem of collective action posed by the back of the Invisible Hand. When we do, our behaviour tends to be, as the egoist's must be, self-defeating. Or so I claim. I must now make the claim good.

4. That egoism is self-defeating in situations involving strategy–payoff conflict may seem evident. For the outcome of egoistic interaction affords each actor a payoff less than he might obtain without the payoff of any other actor being in any way diminished. Everyone could gain; net benefits are possible but not provided. Each, then, does not do as well for himself as possible, and so egoistic behaviour defeats its own end. The egoist aims at maximizing his value and achieves less than the non-egoist who aims instead at Pareto-optimality.

We should not be convinced by this argument. *Each* does not *get* as much as possible, and *all* do not *do* as well for themselves as possible; it does not follow that *each* fails to *do* as well for himself as possible. To show that egoism is self-defeating we must consider, not its overall result, but the situation of the individual who must choose his response to the choices he expects his fellows to make. To consider the plight of egoists solely from the overall or collective standpoint is to commit a version of the fallacy exposed by Olson in his analysis of collective action. Naively, we supposed that a group of individuals maximize their overall net benefit in the same way that a single individual does. Realizing this to be fallacious, we may then suppose that an individual will fail to maximize net benefit in the same way that a group does. But just as what is maximizing from the standpoint of an individual need not be so from the standpoint of a group, so what is self-defeating from the standpoint of a group need not be so from the standpoint of an individual member. And only from this latter standpoint can we show an egoist that his way of acting is self-defeating.

The egoist tells us that we have shown nothing of the kind. He chooses his best response given his expectations of what the others will do. What more can he do for himself? The problem, if indeed it is a problem and not a simple misfortune, is that the choices of others lead to his getting less than he might, given their payoffs. He is the victim of their choices, not his own. They do not seek to victimize him; each in turn simply chooses as best he can for himself. Each is the victim of choices that do not take his benefits and costs into account. But to be victimized is not to engage in self-defeating behaviour. Indeed, were the egoist to complain to his fellows that they did not consider his costs and benefits, they would rightly reply that, were they to take more than their own interests into account, then they would truly be engaged in self-defeating behaviour.

Egoists are defeated by the existence of strategy–payoff conflict. But no individual egoist is defeated. No individual can improve his own lot. The remedy is not for individuals to choose differently, in a non-egoistic way, but rather for them to prevent strategy–payoff conflicts from arising. Those who would otherwise expect to find themselves paying the costs of such conflicts may have good reason to provide for sanctions, through binding agreements or external enforcement, that alter the payoffs so that strategies in Nash-equilibrium lead to a Pareto-optimal outcome. These are the classic devices, proposed by Thomas Hobbes long before the theory of games revealed the precise structure giving rise to conflict. Covenants — but not covenants without the sword, for they are but words of no strength to secure a man-and the sovereign who enforces covenants and structures social institutions to prevent free-riding bring order to the egoists' world.[30] These precautionary devices themselves involve costs that egoists would prefer to avoid, and Hobbes may be accused of failing to give sufficient consideration to these costs,[31] but the world is under no obligation to accommodate itself to all of our preferences. There is nothing self-defeating in the need to cope with structures of interaction that in themselves impede persons seeking the greatest possible realization of their actor-relative values.

The charge that egoism is self-defeating seems to rest on confusion. W e must distinguish between the choices of individuals and the structures within which they choose. The claim that egoism is self-defeating is a claim about the effects of egoistic choices. It can not be supported merely by pointing to the effects of strategy–payoff conflict. These effects determine the possibilities for choice; within these the egoist does the best he can. H e would do better were the possibilities otherwise, were the world a perfect market.

The sensible egoist may of course seek to convince his fellows that egoism is a self-defeating policy. Aware of the costs they inflict on him, he may in his own interest seek to persuade them

[30] See Thomas Hobbes, *Leviathan* (London: 1651), chs. 15, 17.

[31] But Hobbes does give some consideration to this matter. At the end of *Leviathan,* ch. 18, he notes that "the estate of Man can never be without some incommodity or other," and goes on to compare the costs of government with those of civil war and the absence of all authority.

not to impose such costs. He may appeal to the idea of strategy–payoff conflict in the hope of convincing them that, since everyone ends up worse off than need be, their egoism is self-defeating. But his appeal is purely specious, intended to secure for himself the benefits of non-egoistic behaviour by others, while continuing clear-headedly to displace what costs he can upon them — save that he must appear to practice what he preaches to enhance the effectiveness of his preaching. The claim that egoism is self-defeating is, it may now seem, not merely a misunderstanding of the nature of strategy–payoff conflict, but the egoist's deliberate distortion of its real character.

5. The egoist's defence is mistaken. He does not do as well for himself as he could. The reader will no doubt be on his guard when I claim this; perhaps I am the egoist seeking to sucker him or her with my honeyed words. Against this suspicion I can but offer argument. And the key to my argument is this. An egoist will of course maximize actor-relative value whenever he can. Putting to one side those situations in which egoism may fail to offer consistent guidance to choice, let us agree that in each situation the egoist chooses a best response to the choices he expects others to make, and that *in those situations* he can do no better. But his very way of choosing affects the situations in which he may expect to find himself. And the effects are to his disadvantage. The egoist makes the most of his opportunities, but as an egoist he finds those opportunities inferior to those of a non-egoist — not, to be sure, just any non-egoist, but one whom I shall call the cooperator. In making this clear we show the self-defeating character of egoism.

In the Prisoners' Dilemma we may distinguish a cooperative and a non-cooperative strategy for each actor. In the tale of Jack and Zack, the non-cooperative strategy is of course to confess; the cooperative strategy is to remain silent. We may be thankful if prisoners prove to be non-cooperators, if there is no honour among thieves, but in general, and always from the standpoints of those concerned, non-cooperation is costly. Two cooperators will each do better than two non-cooperators. The problem, as we have seen, is that a non-cooperator paired with a cooperator will do better still, and at the cooperator's expense.

Suppose then that an actor is *conditionally* disposed to cooperate in Prisoners' Dilemma situations, and more generally in all situations involving strategy–payoff conflict. She does not unthinkingly opt for a cooperative strategy. Instead, she forms an expectation about the strategy choices of her partners (or opponents) and conforms her own choice to that expectation. She chooses cooperation as a response to expected cooperation, and non-cooperation as a response to expected non-cooperation.

How does her conditionally cooperative disposition affect her payoffs? At first glance it may seem that it must reduce them. If she expects the other to choose a non-cooperative strategy, then she maximizes her expected payoff by her own choice. But if she expects the other actor to cooperate, then she does not maximize her expected payoff and so gains less than were she consistently to choose non-cooperation.

But this argument fails to take into account the disposition of the other actor or actors. Suppose that the other actor is also conditionally disposed to cooperation. Then were she disposed not to cooperate, and could expect him correctly to read her intention, she would expect him also not to cooperate, and so would expect to end up at the mutually disadvantageous outcome of strategies in Nash-equilibrium. But if she is disposed to cooperate, and again expects the other correctly to read her intention, then she expects him to cooperate and so she expects to end up at a mutually advantageous Pareto-optimal outcome. Among conditional cooperators, expectations about others' choices and dispositions to choose oneself are so related that each may benefit from interaction in ways that non-cooperators can not parallel.

The egoist seeks to maximize actor-relative value given his expectations about the strategies others will choose. But their choices, and so his expectations, may be affected by his egoistic, maximizing policy; others, anticipating his choice, respond in a maximizing manner. The cooperator refrains from seeking to maximize value given her expectations about the strategies others will choose. And their choices, and so her expectations, may be affected by her cooperative policy; other cooperators, anticipating her choice, respond in a cooperative manner. And so egoism is self-defeating. Our argument rests on a comparison between the effects of choosing on a maximizing, non-cooperative basis, and the effects of choosing on a conditionally cooperative basis. Although the conditional cooperator refrains from making the most of her opportunities, yet she finds herself with opportunities that the egoist lacks, and so may expect payoffs superior to those that he can attain.

Of course the conditional cooperator may err. She may fail to recognize the willingness of others to cooperate with her, and so treat them as egoists. She may fail to recognize the egoism of others, and, treating them as cooperators, be taken advantage of by them. Unless cooperators are reasonably capable of both identifying one another and singling out non-cooperators, their conditional disposition may prove disadvantageous. This is an empirical matter. However, given the real benefits of cooperation, we should expect would-be conditional cooperators to seek to improve their abilities both to identify the dispositions of those with whom they interact and to make their own disposition known. Although the actual advantageousness of conditional cooperation depends both on these abilities and on the proportion of cooperators in the interacting population, yet the potential advantageousness of the disposition is not empirically based, but reflects the logical structure of interaction. Ideally, an individual whose objective is egoistic, to do as well for himself as possible, must expect to do better, not as an egoist, but as a cooperator.

I claim, then, that given the capacity to choose between egoism and conditional cooperation, and given also sufficient ability to identify the dispositions of others and to make oneself identifiable in turn, a rational person will choose to dispose herself to conditional cooperation, This choice is itself an egoistic one; she maximizes her expected actor-relative value in so choosing among possible dispositions to choose. But its effect is to convert her from an egoist to a cooperator, to a person who, in appropriate circumstances, does not choose egoistically.

Before considering objections to this argument, I should note that it does *not* depend on the supposition that one may expect to find oneself in an indefinite sequence of strategy–payoff conflicts, so that by choosing to cooperate in a particular situation one affects the expectations, and so the choices, of others in subsequent situations. There has been considerable discussion of the importance of reputation in iterated or repeated Prisoners' Dilemmas and in situations in which one benefits from a credible deterrent threat — for example the market-entry situation discussed in Reinhard Selten's "Chain-store Paradox." [32] But our concern is not with reputation or threat. The rationale for choosing conditional cooperation over egoism does not depend on the supposition that one can gain long-term benefit by acquiring a reputation for making cooperative choices. My argument may be applied to a one-shot conflict.[33]

Suppose that each person were to know — weneed not mind how — that once and only once in her life would she face a strategy–payoff conflict. If she could reliably identify the disposition of the other actors in that situation and could expect them to identify hers, then she would have reason to dispose herself to conditional cooperation. For were she able to do so, then, if her partners in the situation were also conditional cooperators, she would do better than were she a non-cooperator. And were her partners non-cooperators, then, since she would so identify them, she would do no worse than had she remained an egoist. Note

[32] See "The Chain-store Paradox," *Theory and Decision* 9 (1978), pp. 127–59.

[33] For the application to deterrence, see my paper "Deterrence, Maximization, and Rationality," to appear in *Ethics 94 (1984),* and in Douglas MacLean (ed.), *The Security Gamble: Deterrence in the Nuclear Age* (Totowa, N.J.: Rowman and Allanheld, in press).

that in this situation, if she finds herself among cooperators, she clearly does *not* maximize actor-relative value, whether in terms of her short-run expectations in the particular situation, or in terms of her long-run expectations for the remainder of her life. If she is genuinely disposed to cooperate, then in appropriate circumstances she does not behave in any way as an egoist.

Let us now turn to objections. It will no doubt be said that, although it may be rational to pretend to be a cooperator, yet it is not rational actually to be one. The rational egoist will not give up his egoism however much he may appear to do so. Now I do not deny that there can be circumstances in which pretence would be the rational policy for a maximizer of actor-relative value. But no argument has been, or can be, given to show that this must always be the case. Perhaps pretence will not work — the detecting capacities of others are too good. Or perhaps the psychological strain of pretence is simply too great. The best way to reap the advantages of cooperation may be to be a genuine cooperator. The honesty that is the best policy may prove to be the honesty that, once adopted, can not be cast aside.

It may then be said that our argument shows that the egoist must recognize the benefits, in appropriate circumstances, of disposing himself to conditional cooperation. But then egoism is not self-defeating. Rather it contains the resources for its own reform. The truest egoism is conditional cooperation. This objection interprets egoism very differently than I have done. In considering whether egoism is self-defeating, as in considering whether it is inconsistent, I have focussed on an egoistic principle for choice among strategies or actions. I have shown that it is self-defeating to be unconditionally disposed to act on such a principle — that is, on a principle satisfying the condition that it take each set of alternative possible strategies for an actor into a sub-set, the members of which maximize some actor-relative measure defined over the original set. I take egoism to be the unconditional disposition to act on such a principle. The person who, for whatever reason, chooses not to act on such a principle, chooses not to be an egoist. Her reason for so choosing may itself be egoistic, as may her choice not to be an egoist. But it is what she chooses, and not why or how she chooses it, that is decisive here; she is not an egoist if she does not choose an egoistic principle for choice.

Egoism does indeed contain the resources for its own reform. The egoist is able to recognize the self-defeating character of his disposition to choose, and so has reason to select an alternative disposition. But the reform ‘that the egoist carries out is not one internal to his original egoistic position. Choosing conditional cooperation is the egoist's last act as an egoist, and in that act the self-defeating character of egoism is affirmed.

6. By limiting the pretension of egoism we enhance the prospect for morality. Were egoistic principles of choice not self-defeating, the moralist would find herself compelled to reject either an actor-relative conception of value or a maximizing conception of rationality. Morality, as I understand it here, provides an *internal* constraint on the straightforward attempt to do as well for oneself as possible. An internal constraint is one that falls between the actor's evaluation and her choice — a constraint, then, on her principle for choice, the function taking her sets of alternative strategies into choice sets. Such a constraint has no claim to the egoist's consideration. He does not consider the imposition of an *external* constraint unjustified, since for him all justification is actor-relative and those imposing the constraint may well find that it promotes their ends. But an external constraint affects either one's range of options, and so the strategies among which one can choose, or the value one may expect from the outcome of some of one's options. An external constraint thus leaves the egoist free to choose on the basis of his maximizing principle, and so leaves his egoism intact. But voluntary adherence to a constraint on maximization — a constraint leaving one's range of options and their values unaffected — is incompatible with egoism. If morality provides such an internal constraint, then a

moral principle for choice must be incompatible with, and so an alternative to, an egoistic principle.

To avoid possible misunderstanding, let me note that in treating morality as an internal constraint on straightforward maximization, I am offering a necessary, but not a sufficient, characterization. Not every conceivable internal constraint would be moral. I can not here offer a full account of what must be added to the idea of an internal constraint to capture the concept of morality; what I shall add will relate morality to cooperation.

In rejecting egoism the moralist has traditionally employed one or both of two lines of attack. The first turns on the egoist's conception of actor-relative value. As I noted in the first part, some, such as G. E. Moore, profess to find this conception self-contradictory, insisting that value must be absolute. Thus they accuse the egoist of maximizing the fulfilment of his interest rather than maximizing what is truly good. His interest may be part of this good, but no more part than the interest of anyone else. Others argue that the egoist mistakes his apparent interest for his true interest and claim that each person's true interest is linked to a transcendent, non-relative value in such a way that the conflict between the apparent interests of individuals is replaced by the harmonization of their true interests. This I believe to be the position Plato advances in the *Republic;* thinkers of this persuasion accuse the egoist of maximizing the fulfilment of apparent good rather than true good.

The second line of attack turns on the egoist's conception of maximizing rationality. Some, such as Kant, would argue that the egoist mistakenly supposes reason to be merely instrumental, determining the means appropriate to given ends, so that he fails to recognize that reason has a practical role quite independent of that set it by interest.[34] Morality, on this view, arises from the

[34] This paragraph is intended as an interpretation of Kant's position as he develops it in the *Groundwork of the Metaphysic of Morals* and the *Critique of Practical Reason,* but it is not my concern here to offer any defence of whatever con-

ascription to practical rationality of the same universality found in theoretical rationality. As theoretical reason discovers descriptive or explanatory laws, so practical reason discovers prescriptive or justificatory laws. The egoist's commitment to maximization is thus rejected as insufficient for the universality inherent in true rationality.

Neither of these lines of attack on egoism seems promising to me. Fortunately, I need not argue that claim here; it would be absurd and presumptuous to think that I could dispose of two of the main traditions of moral thought in a few words. Instead I can bypass them. Allowing the egoist his conception of actor-relative value and of maximizing rationality, I have shown that if he does not give up his egoism in favour of conditional cooperation, then he bars himself from opportunities for advantageous interaction with his fellows. Adherence to a principle for choice that places appropriate constraints on maximizing behaviour may be expected to benefit the adherent. And so I find a place for internal constraint, for a non-maximizing moral principle for choice, by arguing from the egoist's premisses to the rejection of his conclusions.

I can then suggest that my attack on the egoist requires assumptions far weaker than those necessary to the attacks mounted by traditional moralists. Where they assault his position from without, seeking to batter down his premisses, I undermine it from within, showing that the premisses give it no support. I need neither absolute value nor universalized rationality. I can then suggest that moral theorists have resorted to these lines of attack because they have not seen the possibility of defending morality by fighting the egoist on his own ground. And so I can suggest that the appeal of both the Platonic and the Kantian traditions

troversial features the interpretation contains. Whether or not the account is faithful to Kant, it seems to me to raise important questions about the instrumentality of practical reason, and the connection of practical and theoretical reason, that deserve non-Kantian answers.

has depended on the failure to recognize a third way by which the moralist may snatch value and reason from the egoist's grasp.

To add substance to these suggestions, let us see how the failure of egoism offers some positive insight into the place morality can occupy. The egoist fails satisfactorily to resolve strategy–payoff conflicts. I have characterized the disposition needed for such resolution as conditionally cooperative. Cooperation is more than mere coordination; the cooperator selects a course of action promoting mutual benefit and adheres to it against the temptation of individually advantageous defection. Thus cooperation requires a real measure of constraint. If we relate morality to the disposition to cooperate, then moral theory will be, or at least will include, that part of the theory of rational choice that is concerned with the formulation of principles for cooperative interaction. These principles perform the traditional constraining role of morality in such a way that their rationality must be recognized by all those who, sharing the egoist's view of value and reason, realize the self-defeating character of his choices.

In the Prisoners' Dilemma the selection of cooperative strategies is unproblematic. But this is not generally true in strategy–payoff conflict. In the Dilemma there is but one plausible way of cooperating, for there is but one outcome that is both Pareto-optimal and mutually advantageous in comparison with the equilibrium outcome of egoistic behaviour. But in most strategy–payoff conflicts there are many ways of cooperating — many outcomes that are both optimal and superior for each person to what she could expect were each to seek directly to maximize value. Moral principles must enable us to select among these possibilities. If they are to be used effectively as an alternative to general egoism, then they must be reasonably simple and clearly established in accepted social practices and institutions. Cooperation depends on the ability of each cooperator to anticipate the choices of her fellows, and this is possible in general only if those choices reflect widely shared principles.

We should expect moral principles for mutually beneficial cooperation to require such traditional virtues as truth-telling and promise-keeping, as honesty, gratitude, and reciprocal benevolence. But we should not expect all of traditional morality to pass the scrutiny imposed by the cooperative standpoint. In relating morality to rational choice we seek to derive principles independent of any appeal to established practice. We are not concerned with reflective equilibrium.[35] Although it would be surprising, did no commonly recognized moral constraints relate to mutually beneficial cooperation, yet traditional morality as such may be no more than a ragbag of views lacking any single, coherent rationale. My account of morality does not attempt to refine our ordinary views, but rather to provide constraint with a firm foundation in rational choice.

7. The role of moral theory is to provide a reflective and critical standard by which existing moral practices may be assessed and revised. The standard for this reflection and criticism is provided by the individual who asks herself what she may rationally put in the place of an egoistic principle — what principle of choice, if adhered to by everyone, would be acceptable to her. And she must consider not only herself but everyone else; since her adherence to the principle is to be conditional on her expectation of others' adherence, she must expect those others also to be convinced of its acceptability.

Each person prefers to cooperate with others on terms as advantageous to herself as possible. But each must recognize that everyone has this preference. And so no one can expect others, insofar as they are rational, to accept terms of cooperation less advantageous than the least advantageous terms she herself will will accept. The recognition of mutual rationality leads to the requirement that moral principles be mutually acceptable.

[35] For the claim that moral theory is concerned with reflective equilibrium, see *A Theory of Justice*, pp. 20–21, 48–51.

Schematically we may represent the question of determining a mutually acceptable principle for cooperative choice in the following way. Let C be the set of feasible principles, and assume that the egoistic principle e is a member of C. Let each person i define an interval measure $V\ i$representing the value to her of each member of C; thus for any principle p belonging to C, $Vi(p)$ is i's evaluation of p. Let $v(p)$ be the set of all individual evaluations $vi(p)$; we shall call it the *value-representation* of p. And let $V(C)$ be the set of all value-representations $v(p)$; in other words:

$$V(C) = [v(p)\ p \text{ is a member of } C]$$
$$v(p) = [v_i(p) : i \text{ is a person}].$$

Now the acceptability of a principle of cooperative choice depends first on its affording each person greater expected value than does egoism. Beyond this, we may suppose that its acceptability must depend on comparing individual evaluations of the principle with evaluations of its alternatives. In other words, our problem is to select a principle r on the basis of its value-representation $v(Y)$, in relation to the members of the set of value-representations $V(C)$, and the particular requirement that for each individual i, $v_i(Y)$ is greater than $v,(e)$, which we may write as: $v(r)$ is greater than $v(e)$. But this problem is isomorphic with the usual formulation of the bargaining problem in game theory: to determine an outcome, defined in terms of its values, as a point of mutual agreement, given a set of possible outcomes and a fixed outcome representing no-agreement.[36]

In section 2 of the first part I distinguished my view of the relation between rational choice and moral theory from that held by John Rawls. I claimed that moral principles are principles *for* choice, used to select among possible actions or strategies, whereas Rawls treats the principles of justice as objects *of* choice. And I claimed also that moral principles relate to strategic rationality,

[36] For an account of the bargaining problem, see R. D. Luce and H. Raiffa, *Games and Decisions* (New York: Wiley, 1957), pp. 124-26.

and so to situations in which each person chooses on the basis of his expectations of others' choices, whereas Rawls relates the principles of justice to the solution of a problem of parametric rationality in which the circumstances of choice are treated as uncertain but fixed. W e may now see that the distinction between my view and that of Rawls is somewhat more complex than I previously suggested. Moral principles are indeed principles *for* choice, and for strategic choice. They are principles for choice in cooperation. But they are also objects of choice, in that moral principles would be *agreed to* by rational persons, considering possible alternatives to the egoistic principle for situations in which strategy–payoff conflict makes cooperation desirable. They are the principles for choice that would be chosen by all rational persons in situations in which everyone's choosing on the basis of an egoistic principle would be harmful to everyone.

But if moral principles are objects of choice as well as principles for choice, note that they are not the objects of a parametric choice. I have argued that the problem of selecting moral principles is isomorphic to the bargaining problem, and this is one of the central questions in the theory of strategic choice. It seems to me extraordinary that given the role of moral principles in interaction, those theorists who have wanted to relate moral theory to the theory of rational choice, such as John Rawls and John Harsanyi, have not recognized that the theory of bargaining, of strategic agreement, offers the appropriate point of linkage.[37]

In proposing that we consider moral principles as the outcome of a rational bargain, I am not suggesting that morality is a matter of bargaining skills, as these are ordinarily understood. No doubt the principles that would result were actual persons to negotiate among themselves would reflect the differing abilities of the persons and the initial advantages or disadvantages that each would

[37] Rawls does argue for his focus on individual decision rather than bargaining in section 20 of *A Theory* of *Justice.* Harsanyi never considers that ethics and bargaining might be related.

bring to the bargaining table. But although moral principles are of course to be applied by actual persons in their real interactions with their fellows, the bargain by which I suppose them selected is not itself actual. We must abstract from the real situation of actual individuals in two important ways. First, since the principles chosen are to be used as a standard for assessing social practices and institutions, they must be chosen from a position *prior* to the existing social structure. Individuals are to be thought of as choosing principles for their interaction *ex unte,* so that they can not bargain from the particular advantages or disadvantages that the actual workings of society have conferred upon them. Each may bring only his or her natural assets to the bargaining table. And second, the choice of principles is to be determined by the requirement of bargaining theory rather than by actual negotiation among imperfectly rational actors, so that each person is in effect represented at the bargaining table by an ideally rational self, and no question of differential bargaining skills arises. Moral principles are those to which our rational selves would agree, *ex unte,* for the regulation of our cooperative interactions.

Ideally rational selves do not, however, exist behind a Rawlsian veil of ignorance. Rationality here, as throughout my argument, is instrumental. Each person's rational self is fully informed about his or her abilities and interests. Thus the idea of a bargain satisfies the condition, stated but not in my view observed by John Rawls, that moral theory "take seriously the distinction between persons." [38] A rational bargain is rational from the standpoint of each person party to it; the bargain determining moral principles is thus *ex ante* rational for every person. The demand that moral theory be part of the theory of rational choice keeps the individual, not simply as a free and equal moral person,[39] but in all the richness of her talents and interests, her capa-

[38] *A Theory of Justice,* p. 27.

[39] ,The phrase "free and equal moral person" comes from Rawls; it constitutes one of the central themes of "Kantian Constructivism in Moral Theory," *Journal of Philosophy* 77 (1980), pp. 515 - 72.

cities and concerns, her distinctness from her fellows, as the focal point of morality.

8. And so I claim that our argument vindicates morality by appealing directly to each one of us. Not only does each of us do better by disposing himself or herself to conditional cooperation with others, but the terms of that cooperation are determined by an agreement to which each of us is fully party. Each of us may then begin as an egoist, seeking to do as well for himself or herself as possible, and supposing that this maximizing objective must guide each choice, each action. But the argument I have sketched in these lectures should lead us out of egoism; without abandoning the objective of doing as well for oneself as possible, each of us must recognize that the direct translation of that objective into a principle of choice is self-defeating. The morality of conditional cooperation offers the correct translation of the egoist's objective into action.

So what should an egoist do? Why, he should become a cooperator and consent to morality. Here I have only begun the task of illuminating that morality by an appeal to the theory of rational choice. My principle task in these papers has been a preliminary one — to sketch the incompleat egoist so that, in seeing what he lacks, we might better be able to seek out the compleat moralist.