5

Prisoner's Dilemma and Resolute Choice

EDWARD F. MCCLENNEN

In the last twenty-five years a great many very interesting articles have been written about Prisoner's Dilemma. Yet for all that, one has the sense that the discussion has moved less far than one might have hoped, given the rather counterintuitive conclusion with which the dilemma was originally launched. I refer here to the view that for one-shot situations, with choices to be made simultaneously, with no probabilistic (or alternatively, causal) dependence between choices, and conditions of full information, rational players who know each other to be such must play non-cooperatively, despite the fact that this means that they must forgo benefits that would be available to each if only they could manage to cooperate. In the language of the economist, the rational outcome is suboptimal.

There have been some very important developments. Prisoner's Dilemma has proved a powerful diagnostic device for understanding the function of many moral, legal, social, and political institutions. It has also occasioned a much more sophisticated analysis of epistemic conditions and preference patterns under which rational and cooperative choice might coincide.[1] Virtually all of these various explorations, however, have worked from a fixed point of conceptualization according to which a rational agent is one who, on each occasion calling for decision, chooses so as to maximize with respect to an antecedently and exogenously specified preference function, given (again from the perspective of that same occasion for decision) the expected behavior of the other player. That is, the choice situation is conceived as parametrized with respect to two crucial variables: the agent is to maximize a *given* preference function against the *given* (expected) choice of the other player.[2]

On such a view, it remains possible, of course, that the preferences of the two agents are such that each choosing in accordance with the constraints specified above will result in an outcome that is optimal. Alternatively, it may be that each entertains special beliefs, for example, to the effect that their respective choices are causally or probabilistically connected. But it follows from what I want to term the fixed point of the modern theory of rational choice that some story of this sort will have to be told. That is, on this view a potential Prisoner's Dilemma situation will be resolvable only insofar as we can assume something about the particular objects of the players' preferences or the particular content of their beliefs.[3]

I want to try to make a case for thinking about Prisoner's Dilemma in a somewhat different way. Specifically, I want to argue that what it is rational for a player to choose in such a situation is in part a function of a potential in the structure of the game itself for achieving optimality if only there is coordination and not just a function of what would maximize some antecedently specified preference function. Put in slightly different terms, I want to explore the notion that the very preferences a rational agent has in such a situation need to be understood as shaped by the logic of the situation itself.

To pursue such a line of analysis is to travel over ground that has been most ably broken by David Gauthier.[4] Like him, I have become persuaded that there is a need for a reappraisal of the requirements of rational choice as typically presented, a need for a perspective from which cooperation can be understood as arising from the logic of the interaction situation itself. I understand him to advocate that agents choose in the resolute manner I shall argue for in this paper. However, he makes the argument turn on the idea of maximizing expected utility at the level of dispositions to choose instead of at the level of particular choices. I want to suggest instead that the resolute chooser can be interpreted as maximizing utility at the level of the particular choice, but that this utility is contextually dependent on the nature of the interaction situation. My arguments, then, to the extent they succeed, may provide an alternative way to motivate what I take to be a central feature of his position, and to show how it can be defended against the prevailing conception of rationality. The strategy I shall employ is to show that there is an important connection between Prisoner's Dilemma and another type of choice situation about which the theory of rational choice in question has had some interesting things to say. I intend to try to use what it has to say about this other type of choice situation against what it has to say about the rational solution to Prisoner's Dilemma.

THE STANDARD ACCOUNT OF PRISONER'S DILEMMA

To fix on a simple example, consider the Ring of Gyges story to be found in Plato's *Republic*. Gyges had a ring which, when he turned it on his finger, made him invisible. This gave him great power: with its aid, he was able to seduce the queen, kill the king, and seize the throne for himself.

In the original story, of course, only Gyges had such a ring. Suppose, however, each of a number of persons who interact with one another pos-

sesses such a ring. If all use their rings, the result is likely to be mutually disadvantageous. Whatever a given person gains by using her ring will be more than offset by the losses incurred by others using theirs against her. Each, then, will stand to gain from all not using their rings. But, of course, for any given player use dominates non-use: no matter what others do, a given player will do better to use her ring. This situation, simplified to the case of two persons, serves plausibly as an interpretation of the usual Prisoner's Dilemma matrix (with the left-hand and right-hand numbers in each box giving the preference ordering for Row player and Column player respectively—a larger number meaning more preferred):

		Column	
		NO-U	*U*
Row	*NO-U*	3, 3	1, 4
	U	4, 1	2, 2

WHO IS RESPONSIBLE FOR THE PROBLEM HERE?

For diagnostic purposes, the game given above has complicating features. *U* dominates *NO-U*, for each player, and on the assumption that both players are able to correctly anticipate what the other player will do, only the pair of strategies (*U, U*) is such that each member of the pair maximizes the antecedently given preferences of that player, given that the other player chooses the other member of the pair. That is, only (*U, U*) is in equilibrium with respect to the specified preferences of each. But independent arguments can be brought in support of a choice of *U*. *U* is the safest strategy for each: it maximizes a player's security level. Each player, then, can plead a willingness to be cooperative, were it not for the need to secure against non-cooperation by the other player.

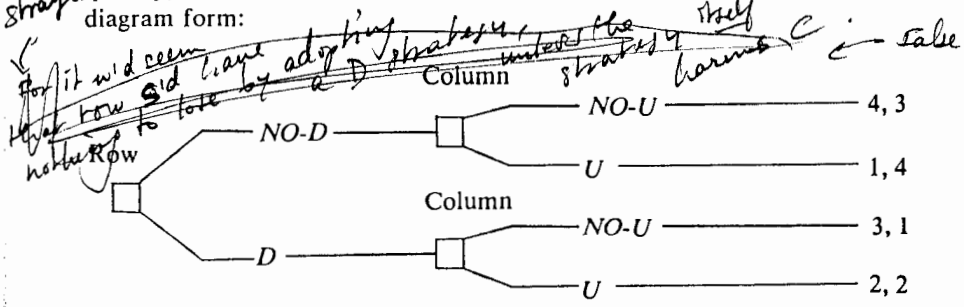
We can obtain a version of this game without this complication by changing the payoff matrix and description of strategies in the following fashion:

		Column	
		NO-U	*U*
Row	*NO-D*	4, 3	1, 4
	D	3, 1	2, 2

To give this an interpretation, suppose that only Column has a ring, while Row has but a partial defense that can be marshalled against it. Let *D* be Row's option of providing such a defense and *NO-D* the option of not so providing. As before, let *U* be Column's option of using the ring, with *NO-U* the option of not using it. The situation is one in which both players would be better off with the outcome of *NO-D/NO-U* than with the outcome of *D/U*, but Column's dominant strategy is *U*, and Row's rational choice might plausibly be taken to be *D*. On the usual account, the outcome of this game, then, when played between completely rational agents who know each other to be such and who are able to correctly anticipate each other's choices, will be *U/D*, to the mutual disadvantage of both.

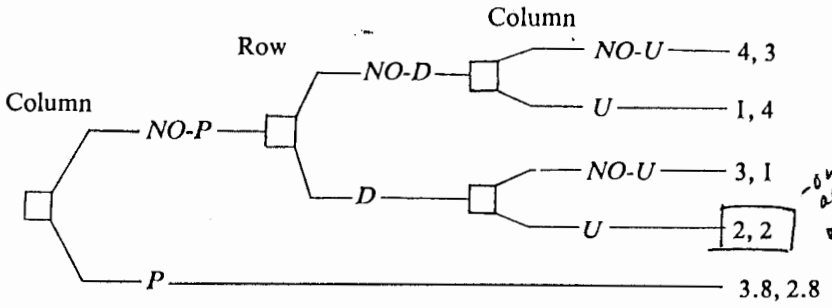
In this game, Column cannot plead that Row's disposition to non-cooperation requires a security-oriented response of *U*. Row's maximizing response to a choice of *NO-U* by column is *NO-D*, not *D*. Under conditions of full information and correct anticipation by each of what the other will do, it is impossible to rationalize a choice of *D* by Row, except on the hypothesis that Column must choose a dominant over a dominated strategy or, more generally, must choose so as to maximize her antecedently specified preferences, given what she expects the other player to do. Thus, it is Column's own maximizing disposition so characterized that sets the problem for Column.

This point can be driven home even more clearly by considering one more alteration in the game. Let us interpret it as a game in which choices are to be made in sequence, with Row going first and Column choosing subsequently, after being informed what choice Row has made. In standard tree-diagram form:



Introducing this particular sequencing of choices changes nothing for a theory that requires rational choices to be in equilibrium with respect to antecedently specified preferences. Column's rational choice, when it comes her turn, will be *U*, regardless of which branch of the tree Row choice places her on, and thus row player must expect column player to select *U* over *NO-U*. But then, the rational choice for row player will be *D*. Consequently, the outcome of rational interaction will be *D/U*. Yet if Row player had reason to suppose that Column player would choose *NO-U*, her best opening move would be *NO-D*, and the outcome would be *NO-D/NO-U*. Once again, then, Column's problem turns upon her own disposition to choose, not the disposition of Row.

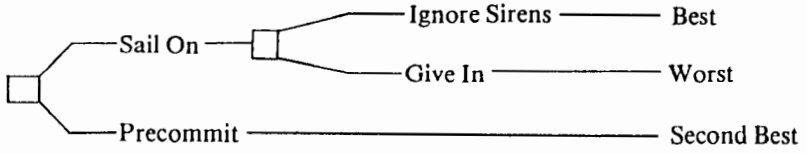
There is one condition under which players do not have to forgo the mutual benefits of cooperation that are possible in this game. Let us suppose that prior to making a choice in the game specified above, there is some way Column can precommit to a strategy of *NO-U*. Of course, such a precommitment device will typically require the expenditure of some resources. If this is so, and if the players avail themselves of such a device, the payoffs for each will be somewhat less than those associated with their playing the combination *NO-U* and *NO-D*, say (3.8, 2.8) instead of (4, 3). In tree-diagram form, precommitment (*P*) can be represented as a strategy that Column can choose, prior to Row choosing:



Given the costs of precommitment, Column will prefer the outcome denoted by (4,3) to the outcome of precommitment. But, on the view in question, for the reasons already rehearsed, this outcome is not accessible. And once again, Column cannot blame the problem on the dispositions of Row. Row would be more than willing to play *NO-D*, were it not for the dispositions of Column. Thus, Column's quarrel is with herself. Or should we say, with her own future self? The outcome that Column would most prefer cannot be obtained because, at some subsequent point in time, she herself would prefer to choose *U* over *NO-U*.

ULYSSES AND THE SIRENS

The analysis has led us from one Greek myth to another. The mutual Ring of Gyges problem can be reinterpreted as a version of the problem of Ulysses and the Sirens.[5] As Ulysses approaches the island of the Sirens, he has no desire to be detained by them; but if he acts on his present preferences (to get home as quickly and as inexpensively as possible), he faces a problem. He is informed that once he hears the Sirens, he will want to follow them. Since *here, now* he does not desire to have this happen, he precommits. He buys wax to stop up the ears of his sailors, good strong hemp with which to have himself bound to the mast, and (what is perhaps most costly of all) arranges for his first-mate to act as his agent. In tree-diagram form, Ulysses's problem looks like this:



FROM ULYSSES TO ALLAIS

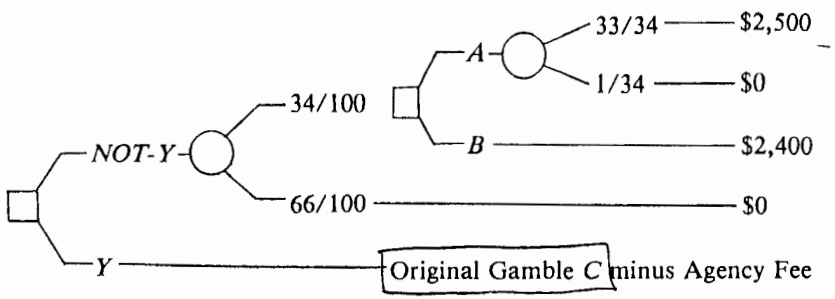
It may be objected that Ulysses faces a very different problem from the one faced by the players in any of the Prisoners' Dilemma games we have explored. The argument is that the latter deal with problems faced by fully rational agents, while the former describes a case in which the self is thought to be temporarily overcome by some irrational (or non-rational) force. But this feature of the story is not essential. We have only to suppose that Ulysses realizes that he is in a situation in which he can predict that his preferences will undergo a specific change.

By way of illustration, consider a sequential version of a paradox that has attracted a great deal of attention in the literature: the Allais paradox. In the original (non-sequential) version, the agent must choose between two gambles, *A* and *B*, and again between two other gambles, *C* and *D*, with the following schedule of payoffs and probabilities:

	A	(\$2,500, 33/34; \$0, 1/34)
Ⓑ		(\$2,400, 1.0)
	C	(\$2,500, 33/100; \$0, 67/100)
	D	(\$2,400, 34/100; \$0, 66/100)

Studies have shown that many people prefer *B* to *A*, but prefer *C* to *D*. The reason often given is that *B* is to be preferred to *A* because *B* involves no risk; while *C* is to be preferred to *D*, because, both involving substantial risk, *C* has the larger possible payoff. Such a preference pattern is characterized as "paradoxical" on the grounds that, as natural as it may seem to many decision-makers, it violates a fundamental axiom of rational choice, the independence axiom.[6]

Here is the sequential version of this problem:



Let *Y* be the strategy of, say, paying an agent a small fee to execute a choice of *A* over *B* in the event the second choice node is reached[7]. Suppose now the player considers the option of *NOT-Y* and contemplates choosing *A* over *B*, if the opportunity presents itself. From the perspective of the first choice node, such a plan is equivalent to opting for *C*, since it offers the prospect of getting \$2,500 with probability (34/100 × 33/34) = 33/100 and \$0 with probability (66/100 + (34/100 × 1/34)) = 67/100. Since *C* is presumably preferred to *Y* (it promises the same schedule of payoffs and odds but without the agency fee), one might suppose that the agent will be disposed to reject *Y* in favor of *NOT-Y* and *A*. But the agent who has the preference patterns described above must reckon with the following consideration: in the event the second choice node is reached, the choice is between *A* and *B*, and by hypothesis, she prefers *B* to *A*. On the usual account, then, if she chooses *NOT-Y*, she must expect that, if the opportunity arises, she will choose *B* rather than *A*. But given this prediction, a choice of *NOT-Y* is equivalent to *D*: it offers the prospect of \$2,400 with probability 34/100, and the prospect of \$0 with probability 66/100. Since, then, by hypothesis, she prefers *Y* to *D*, her rational choice will be *Y*.

Agency fee must be smaller than \$100.

THE POLITICAL ECONOMY OF INTERACTION BETWEEN THE PRESENT AND THE FUTURE SELF

On the view in question, as the above example is designed to illustrate, the present self must assume that its future self will maximize with respect to the antecedently specified preferences it has for the options available to it. Thus, the task confronting the present self is to identify that option presently available to it that will maximize its own antecedently defined preferences, given such maximizing behavior by its future self. The present self who adopts such a stance is characterized as a sophisticated chooser.[8]

Nothing in this account requires of such selves that they coordinate their actions. The present self, to be sure, adjusts its choice in the light of what it expects the future self would choose in response to its own choice—what it would choose from among various possible choice sets (that the present self has the power to make available to the future self). The future self must live with the choice the former self makes, but it has its own agenda of preferences, and when it comes time for it to choose, it maximizes with respect to those preferences. Both, then, are to be understood as maximizers from their own perspective, subject in each case to the constraints that maximizing behavior by the other places on them. If one thinks of the present self as principal, nothing requires the future self to think of itself as an agent, as beholden to some plan of action initiated by the former self.

One must note, of course, that the future self *may* have an interest in coordinating with its past self. Nothing in the analysis so far precludes a preference system for the future self that is responsive to the idea of itself as beholden to the projects of the past self. Rationality, on the view in question, does not rule out such an attitude. But on such a view the attitude must be accounted for *exogenously,* that is, it must be built into the antecedently specified preference of the agent.

The Ulysses-Allais problems of preference change and the various sequential versions of Prisoner's Dilemma are thus seen to be suggestively linked together. In each case, rationality is spelled out in terms of the notion of sophisticated choice. The self's task is simply to adjust its choice in order to maximize (in terms of antecedently specified preferences) against what it takes to be its own future behavior.

Sophisticated choosers do the very best they can, given the constraints imposed by the need to adjust present choice to future given behavior. This is what Ulysses does and what most of us do.[9] But, in the language of the economist, such an approach to choice involves a retreat to second-best. As the various tree-diagrams we have been exploring make clear, sophisticated choosers will typically forgo certain possible gains. If Ulysses manages to save himself from the cursed isle by means of wax, hemp, and agency arrangements, still he could have done even better if he had simply *resolved* to sail right by the island and pay the singing sisters no mind. But so, also, if rational players who know each other to be such do well enough by devising various precommitment schemes or other devices to provide one another with incentives to cooperate, they could do better still by simply resolving to so cooperate.

The connection, however, goes deeper yet. The view that has emerged in the literature on rational dynamic choice is that preference shifts of the Ulysses-Allais type are irrational. Indeed, it has recently been argued that the case for certain of the axioms of utility theory itself rests on the consideration that those whose behavior fails to conform to these axioms will face preference shifts of the type just discussed, shifts that will require resort to second-best strategies, to precommitment and other forms of sophisticated choice.[10] And this, in turn, means that agents whose preferences undergo such shifts will do less well than those whose preferences are not subject to such shifts.

But here the story takes a surprising turn. It is the very theory of rational choice we have been examining that insists that rational agents will be unable to carry through a resolve to act cooperatively in a single-play Prisoner's Dilemma game and that the only recourse will be to various precommitment strategies. But this implies that the preferences of those who act on this theory of rationality are subject to shifts. When faced with Prisoner's Dilemma, each agent wants, *ex ante,* to cooperate, but *ex post,* when called upon to cooperate, each will be disposed to defect. The theory insists, then, upon the reasonableness of preference changes in the one case that are quite analogous to the very preference changes it takes to be irrational in the other case.

RESOLUTE CHOICE AND PRISONER'S DILEMMA

What has gone wrong here? What the theory of rationality in question fails to make room for is the concept of a plan to which the agent might commit herself, not because the plan in question is consonant with some exogenously given set of preferences for various abstractly conceived consequences, but because such a plan resolves a problem posed by the specific interactive logic of the situation confronting the agent. It is a conception of rationality that cannot envision preferences as taking their shape in part from a sense of the limits of strategic interaction because it can only think in terms of what strategy is required by exogenously specified preferences.

Ulysses can solve his problem, I want to suggest, by resolving to sail right by the island and then choosing, when the occasion presents itself, to act on that resolve. Rational agents who know each other to be such can similarly solve the problem they face in sequential versions of Prisoner's Dilemma. The agent who is to choose second has only to resolve to choose in a coordinative fashion and then act on that resolve. If this is the rational approach, the player who is to go first has nothing to fear (at least under conditions of perfect information and a mutual sense of each other's rationality). But given this, and returning to the first two versions of Prisoner's Dilemma, the conclusion is irresistible: rational players who know each other to be such should have no trouble coordinating in these games. Those from whom a resolve is needed (Column in the second version, both players in the original version) have only to so resolve and act accordingly.

Some will argue that this conclusion carries its own measure of paradox. It would seem that what is recommended is that an agent evaluate the situation from some other standpoint than the preferences she has at the time a decision is required, and one is tempted to question whether any sense can be made of that. Moreover, the suggestion that a rational agent will, at a certain point in a decision tree, choose other than what she prefers at that point plays hob with the whole notion of revealed preference: if the agent chooses *A* rather than *B,* then there is a perfectly appropriate sense in which what the agent really prefers (the preference revealed by choice) is *A* rather than *B*.

The point is well taken, but it hits wide of the mark. The claim that rationality calls for resolute choice need not be read as a claim that the agent must deny her *ex post* preferences. The agent who resolves to coordinate and then acts on that resolve is to be understood as viewing *ex post* choice differently than one who does not so resolve. The suggestion is that rational agents who have an *ex ante* preference for cooperation will prefer, *ex post,* to act cooperatively. For such agents, the *ex post* situation is different from what it would have been if there had been no *ex ante* resolve. In this respect, then, we can claim that resolute players choose strategies that are in equilibrium with respect to what might be termed their considered preferences, that is, the preferences they have for various options, given their understanding of the logic of the interactive situation they face and the need for coordination if an optimal outcome is to be achieved.

It does not follow, of course, that such an agent conforms to all the canons of the view of rationality in question. Some have recently written as if there is leverage against the rationality of resolute choice to be found in an even more general principle of rational choice. Consider the person who, facing Prisoner's Dilemma, resolves to cooperate and then acts in good faith when the time comes, *ex post,* to choose. Imagine now that this agent also insists that if there had been no mutual sense of a need to coordinate, but the situation was otherwise the same, she would have seized the advantage. The suggestion is that such behavior violates a fundamental standard for normative behavior, a consequentialist principle to the effect that if each of two situations yields precisely the same sets of alternative consequences, then a rational agent must choose the same way in the two situations.[11]

Against this it must be insisted that such a principle, if it is to have the import intended, must construe the concept of consequences in such a narrow way that a failure to act on one's resolve cannot count as a relevant consequence. Such a principle is indeed one that the resolute chooser violates, but so construed the principle is surely doubtful. On the other hand, if consequences are construed broadly enough, then there is no need to suppose that the resolute chooser violates the principle in question.

NOTES

Acknowledgements. Versions of this paper were read to the members of the Philosophy Group at Carnegie-Mellon, The Center for Philosophy and Public Policy at the University of Maryland, the Department of Philosophy at Dalhousie University, the Department of Philosophy at Washington University, and the Council For Philosophical Studies Public Choice Institute, held at Dalhousie University in the Summer of 1984. I am indebted all of these audiences for helpful comments, but in particular to my colleague Teddy Siedenfeld, to Peter Hammond of the Department of Economics, Stanford University, and to Mark Machina of the Department of Economics, University of California, San Diego. Founding support for the project with which this paper is connected was provided by the National Science Foundation, under Grant No. SES-8210730.

1. These developments are well documented in the other papers contained in the present volume, particularly the exploration of epistemic conditions. Barry and Hardin (1982) reprint some of the most important contributions to this literature, particularly those concerned with the relevance of Prisoner's Dilemma to the rationale for various institutions.
2. The *locus classicus* for this is the treatment of Prisoner's Dilemma in Chapter 5, Luce and Raiffa (1957).
3. Traditionally, one such line of exploration is to imagine that the players have moral commitments towards one another. But once these commitments are incorporated into the antecedently specified preferences of each player for the various outcomes, the resultant preference matrix is not necessarily a Prisoner's Dilemma matrix. Alternatively, one can explore various special beliefs the agent might have. Consideration of beliefs about probabilistic dependency between the choices of the two agents leads, of course, by one route, to the framing of epistemic conditions under which coordination is rational and, by a somewhat different route, into the impass of Newcomb's Problem and the whole issue of causal versus probabilistic dependency. See, in particular, the papers in Sections III and IV of the present volume.
4. Of particular relevance here are Gauthier (1975) and "Maximization Constrained: The Rationality of Cooperation," in the present volume, pp. 75–93.
5. I must acknowledge a debt here, of course, to Jon Elster, who has written so perceptively concerning the relevance of the problem of Ulysses and the Sirens to the subject of rational choice. Space limitations preclude my exploring fully the respects in which my analysis relates to his. Very roughly speaking, however, he has been preoccupied somewhat more with the concept of rationality as an explanatory (as distinct from a normative) concept. The particular use to which I have put the analogy between Prisoners' Dilemma and Ulysses and the Sirens is one that, I suspect, he would not accept. See in particular Elster (1979) and (1982).
6. The paradox in question is named after Maurice Allais, who was the first to propose the relevance of this particular type of preference pattern to the foundations of expected utility theory. Allais's contribution, which dates from the early fifties, together with some important papers by others and a very extensive bibliography, is to be found in Allais and Hagen (1979). The suggestion that Allais-type preferences imply preference changes in a dynamic choice context is raised by Raiffa in the course of an extremely lucid and interesting discussion of the Allais Paradox in Raiffa (1968, pp. 80–86). See, however, McClennen (1983), for some second thoughts about the success of Raiffa's argument, an overview of the issue, and a bibliography of the relevant literature.
7. Following in the tradition of the story of Ulysses and the Sirens, one of the tasks, it is alleged, that such an agent will have to perform is refusing to listen to any revised order phoned in by the principal if and when the second choice opportunity occurs. My colleague Teddy Seidenfeld is always eager to offer his services as such an agent, but, of course, he insists upon a fee!
8. There is an extensive literature on changing preferences, sophisticated choice as a strategy for dealing with it, and, more generally, what has come to be known as the problem of "dynamic consistency" posed by such preference changes. For two recent surveys, see Hammond (1976) and Yaari (1977).
9. For a recent survey, see Thaler and Shefrin (1981). See also Schelling (1984), Essays 2, 3, and 4 in particular.
10. See Hammond (1976), in particular, and also his (1982a). A very formal and powerful statement of the potential relevance of preference change and dynamic inconsistency to the foundations of utility theory is to be found in his (1982b).
11. The argument is due to P. Hammond. See the last two papers cited in the previous note.

III

EVIDENTIAL VERSUS CAUSAL DECISION THEORY