CONSTITUTIONAL POLITICAL ECONOMY, VOL. 4, NO. 2, 1993

RATIONALITY, CONSTITUTIONS, AND THE ETHICS OF RULES

Edward F. McClennen*

It seems clear that inidividuals stand to mutually benefit, in a wide variety of situations, from structuring their interactions in terms of constitutional practices. But a commitment to treat the rules defining such practices as setting real constraints on choice—a commitment to what could be characterized as an "ethics of rules"—is hard to reconcile with the standard, consequentialist theory of rational choice, which requires, in effect, that individuals regard any rule as providing no more than a "maxim" for choice. Such a theory of rational choice, then, constrains individuals to settle for a second best outcome in which choice is aligned with practice rules by means of strategies of precommitment and threats. The outcome is second best because such methods yield only partial alignment, and involve the expenditure of scarce resources, as well as a sacrifice of flexibility and freedom. To say this, however, is to identify the corresponding theory of rational choice as having *consequentially* unacceptable implications. A modified theory of choice is presented, which is still consequentially oriented, but which assesses consequences in a more holistic manner. It is then argued that this modified theory can provide a rational choice grounding for the needed "ethics of rules."

Introduction

In the study of constitutional political economy, it is customary to stress the importance of rules. This is not simply because the basic political and economic institutions which are the focus of any such study can be interpreted as systems of rules. Beyond that analytic point, there is the consideration that such rules significantly structure—for better or for worse (as measured by standard political and economic criteria)—the way in which persons interact with one another in both the political and the economic sphere. This suggests the importance—for both positive and normative theory—of the comparative study of alternative systems of political and economic rules. But rules can matter in another, non-comparative sense: those who can structure their interactions in terms of rules may be able to achieve things that cannot be

*Regents Scholar in Moral and Social Philosophy, Bowling Green State University, Bowling Green, OH 43403-0222. The author would like to thank three anonymous referees for their very helpful comments and suggestions.

achieved otherwise. In a wide variety of different political and economic settings, a case can be made for rule governed behavior as such. This is implicit in Hobbes' famous remark that when men have no other security than what their own strength and their own invention shall furnish them, the "life of man [is] solitary, poor, nasty, brutish, and short"; and Hume makes the same point, less apocalyptically, when he argues the virtues of a stable set of property rules.[1]

Reflection suggests, however, an even deeper point. When individuals are unable to voluntarily accept the discipline of acting subject to the constraints of rules, a remedy of a sort can be found in one or another kind of enforcement procedure. But it is an imperfect remedy. Enforcement devices typically ensure no more than a partial alignment with rule governed behavior. And what is achieved, at any rate, is second best. This is because enforcement procedures are costly: they consume scarce resources, the requisite surveillance is destructive of personal privacy, and sanctions, when they must be applied, deprive persons of freedom. The conclusion, then, is that truly effective and efficiently organized institutions require an "ethical" climate in which interacting persons voluntarily respect constraints on the manner in which they pursue their interests—require, that is, the acceptance by persons of an "ethics of rules."

Just what this involves can be clarified by reference to Rawls' (1955) distinction between two kinds of rule. Some rules—those which Rawls characterizes as *maxims*—serve merely to summarize past findings concerning the application of some much more general choice-supporting consideration to specific cases. Rules of this sort presuppose both choice-supporting considerations and cases that can be described independently of making any reference to such rules. Correspondingly, exceptions to such rules can be made by direct appeal to the choice-supporting considerations. *Practice* rules have a very different status. While a practice can be defended by appeal to various considerations, the practice itself is prior to the cases to which it is to apply, and serves as a constraint upon choice: that is, a practice determines the range of

1 See Hobbes (1651), part I, chap. 13, and Hume (1888), book III, part II. sections II and III. For a contemporary statement see, in particular, Brennan and Buchanan (1985). In Buchanan (1991a) the argument is even extended to the need for a "work ethic."

appropriate choices in any situation covered by the practice. In particular, taking exception to a rule defining a practice is not something that can be justified by direct appeal to whatever underlying considerations gave rise to the practice. More specifically, it is essential to the idea of a practice that those participating abdicate the right to make decisions case by case by direct appeal to such underlying considerations. It is consistent with this, of course, that the persons who participate in a practice can debate the merits of its rules, and that there are various ways in which to effect changes in those rules. Here, however, no less than at the level of the rules of a "first order" practice, one can encounter practice constrained choice. That is, the process whereby the rules of a practice get modified is itself typically rule governed.

Recast in these terms, the point would be that at virtually every level of political and economic interaction, individuals must accept the constraints of practices, if they are to fully realize the mutual gains that cooperation makes possible. But here one encounters a serious problem. There is a significant tension between this conclusion and the standard assumption of virtually all studies in political economy, that persons do, or at least should, choose so as to promote their own personally defined interests—that is, so as to maximize subjective expected-utility. The problem is simply that, in any given situation, an individual may have a utility maximizing reason to either deliberately violate the rule, or invest time and effort in altering the rule in a fashion that undercuts the original purpose of the practice. It is more than bare possibility, however, that works against a rational commitment to practices. What stands in its way appears to be what stands in the way of a rational commitment to contribute to a public good: what benefits the individual is not that she act in accordance with the rules of the practice, but that those others with whom she interacts do so. The good consequences for any given agent flow, not from accepting the practice as constraining her own action—that is *cost*—but from others accepting the practice as a constraint. Under these circumstances, so the standard argument goes, each person will be rationally motivated to treat the practice as no more than a guideline or maxim for her own consequentially oriented, interested choice, and, even then, as no more than an arrangement that can be changed whenever that serves her interests.

Must one conclude then, that because individuals are constrained by their own rational dispositions, they have no choice but to make use

of second best enforcement mechanisms? At the very least, this points to a diagnostically promising way to interpret the problem. The suggestion is that each individual faces not simply a problem of interpersonal conflict, but a problem of *intra*personal conflict. That is, the individual is to be conceived as divided, in a sense, against herself, and this, moreover, as a result (it would seem) of reasoning, both naturally and plausibly, by reference to her preferences with regard to the *consequences* of her own choices.[2]

What follows is a systematic exploration of this way of interpreting the problem of a rational commitment to rules. The guiding hypothesis is that once this type of intrapersonal problem is carefully analyzed, it points the way to its own solution, and that solution in turn points the way to the solution of the interpersonal problem as well. A warning here is in order. The analysis turns upon an examination of some rather abstract models of rational choice—models that may seem far removed from those of constitutional choice. The reader's indulgence is requested. The constitutional problem is rooted in a way of thinking about rational choice *in general*, a way that is so deeply ingrained in our thinking as to virtually escape attention altogether. The point of focussing upon very simply models, then, is that one may be able to bring the limitations of this way of thinking into much sharper relief.

The analysis commences, in Section I, with an exploration of some standard models of intrapersonal choice—models which suggest some inherent limitations to consequential reasoning. Section II develops the thesis that it is not consequentialism *as such*, but only an incremental version of consequentialism, that generates the problem. This paves the way for a presentation, in Section III, of an alternative, and more holistic way of thinking about consequences. Section IV argues for the superiority of this alternative conception. Section V returns to the case of *inter*personal choice, and argues that the results for intrapersonal choice can be extended to the logically special case of interdependent choice under (ideal) conditions of mutual rationality and common knowledge. Section VI then explores the prospects of extending this argument to more "realistic" models of bilateral and n-personal interaction. In Section VII, the results of the previous sections are explicitly brought to bear on the rationality, in a wide range of situations, of accepting an ethics of rules—of accepting the constraints of practices.

2 Such an interpretation is central, for example, to Vanberg and Buchanan (1990).

Section VIII then brings closure to the argument as a whole by explicitly tracing the implications of this analysis for a theory of constitutional choice.

I. Intrapersonal Conflicts

The formal analysis of *intra*personal conflict situations dates from a seminal article by the economist Strotz (1956), who takes as his epigram the story of Ulysses and the Sirens. As Ulysses approaches the island of the Sirens, he has no desire to be detained by them; but if he simply acts on his present preferences (to get home as quickly and as inexpensively as possible) he faces a problem, for once he hears the Sirens, he will have to follow them (or at least so he believes). Ulysses' problem, then, is that he is divided against himself: there is a conflict between what his present self prefers and what (he projects) his future self will prefer.[3]

3 Ulysses' problem, it should be noted, is subject to two different interpretations. On the first, the Sirens and their song is a metaphor for a situation in which an agent anticipates that his will-power will be quite literally overwhelmed by some external power (the Sirens and their song). Faced with this, Ulysses reasonably takes precaution, and has himself tied to the mast. When the story is interpreted in this way, it connects naturally with, for example, the problem posed for a rational agent by a drug that is physically addictive. On a second interpretation, the story is simply a metaphor for a situation in which an agent projects that his preferences will change over time or with some change in perspective that a temporal shift can imply. On this interpretation, the agent is presumed to be disposed to deliberate and decide *incrementally* rather than *globally*—to reassess options at each choice point in time in terms of interests (short-range or long-range) *as they are perceived from that point in time*, without regard to whatever interests were previously projected with respect to those same choices, and which originally formed the basis for a choice of a plan. On this interpretation, the problem that Ulysses faces is that while he is prepared to adopt a certain plan, he is also disposed, once the plan is adopted, to depart from it, even though, as the story is told, he comes into possession of no new information about his choice situation. In what is to follow, the focus is exclusively on "intrapersonal struggles" of the *second*, rather than the first, kind. That is, the concern is with agents who are potentially disadvantaged by what might fairly be characterized as the "Siren's song" of incremental reasoning, rather than those whose deliberative powers are overwhelmed by external forces.

The logic of this sort of situation can be captured by appeal to a very simple abstract model, involving an agent who must make a pair of choices in sequence, and who faces a *potential* problem of a preference shift (why the qualifier "potential" here will become clear shortly). Suppose that some intelligible story can be told to the effect that outcome o_1 is preferred to o_3 and o_3 is preferred to o_2 at time t_0, but o_2 is preferred to o_1 at time t_1.[4] Now, let the paths by which these outcomes can be reached be as follows:

Suppose the agent considers the plan that calls for her to move to the second choice node and then choose o_1 over o_2. Call this plan a_1-a_3. Since the outcome of this plan, o_1, is preferred at t_0 to the outcome of plan a_2, namely o_3, one might suppose that the agent will be disposed to pursue the former rather than the latter. Unfortunately, according to Strotz, plan a_1-a_3 is dynamically inconsistent. The agent who has the preferences described would, it would seem, upon arriving at choice point 2, choose a_4 over a_3, since, by hypothesis, o_2 is then preferred to o_1. It seems plausible to suppose, then, that were the agent to move to the second choice point, she would end up executing plan a_1-a_4 rather than a_1-a_3. To adopt such a plan as a_1-a_3, and then proceed to depart from it, is to choose in a *myopic* manner. Myopia is here to be understood in a very specific sense: at the initial stage of planning, one does not take into account one's own presumably predictable future

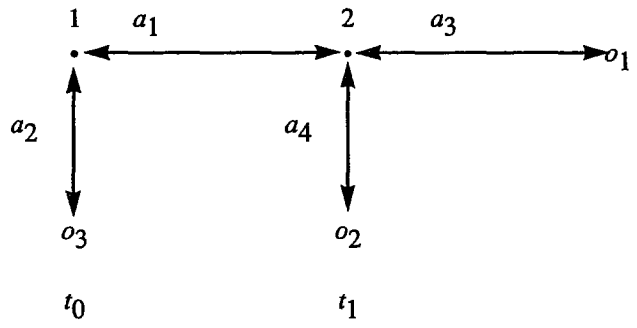

Figure 1. A Simple Sequential Choice Problem

4 See Strotz (1956), Hammond (1976, 1977), Yaari (1977), McClennen (1990a), and Ainslie (1993), for a sense of the wide range of stories that can be told here.

choice behavior.[5] Ulysses, for example, would have been myopic if, taking no precaution, he had planned to just sail right by the Sirens, but, then, upon hearing their song, had abandoned that plan.

Strotz goes on to suggest that rational agents can escape from myopia by choosing in a *sophisticated* manner. They do this by adhering to the principle of *consistent planning*, which requires that plans be tailored to a projection of what one will do in the future.[6] In the problem given in *Figure 1*, a sophisticated chooser must recognize that plan $a_1 - a_3$ is not a feasible plan, and thus that outcome o_1 cannot be secured. Given this, and given her preferences at the first choice node, a sophisticated chooser will select a_2 outright, and thereby realize o_3. There is more than one way to be sophisticated, however. Consider the following interpretation of plan a_2 and associated outcome o_3 : let plan a_2 involve hiring (for a fee) an agent who will execute written instructions, given in advance and irrevocably, to choose o_1 rather than o_2 at the second choice point. By so interpreting plan a_2, a strategy of precommitment can be seen to satisfy the requirements of sophisticated choice. Thus Ulysses has himself tied to the mast, and arranges for his mates to steadfastly sail by the Sirens, regardless of any new orders that he predictably will issue. Ulysses is a sophisticated chooser, and he manifests that sophistication by choosing to precommit.

5 Such persons are not necessarily myopic in the more traditional sense of failing to make long-range plans. The person who saves for next Christmas is projecting into the future, and making a decision now with an eye to what she now prefers to have happen at some future point in time, namely, that she have enough money in mid-December to purchase Christmas presents. Still, she can be myopic in the more particular sense here under consideration, if she fails to reckon with a shift in her own preferences as time unfolds, fails to reckon, that is, with a subsequent desire to take the money out of savings and satisfy some other, more temporary, desire.

6 This way of thinking about the problem of dynamic inconsistency is characteristic of the models to be found in the original work of Strotz (1956), Pollack (1968), Peleg and Yaari (1973), and Hammond (1976, 1977). Yaari (1977) is especially illuminating in this regard. Moreover, with some qualification, dynamic inconsistency is subjected to a similar diagnosis in Schelling (1978) and in Elster (1979). The needed qualification is that Schelling acknowledges the possibility, and Elster explores at considerable length various forms, of "character" training or development. For a fuller discussion of Elster in particular, see McClennen (1990a: chap.13,7). There is also a very interesting and relevant discussion to be found in Buchanan (1979). None of these, however, address directly the issue of the possibility of, and rationale for, rethinking the disposition to deliberate incrementally.

Consider now what is presupposed in this account of *both* myopic and sophisticated choice. Notice, first, that there is an explicit appeal to the principle that a person should choose so as to maximize with respect to preferences for consequences. That is, a person's assessment of the alternatives available is presumed to turn on preferences for the outcomes realizable by her choices. Let this be characterized as the *principle of consequences*. This principle by itself does not suffice to determine that were the person to find herself at the second choice point (in *Figure 1*), the rational choice for her would be a_4 rather than a_3. In fact, within the context of this example, an appeal to preferences for *outcomes* yields conflicting recommendations, since, by hypothesis, while o_2 is preferred to o_1 at t_1, o_1 is preferred to o_2 at t_0.

What is needed, in addition, is an assumption to the effect that it is only preferences for outcomes *still available at a given choice point* that are relevant to choice at that point. Such an assumption is typically secured by appeal to a *separability* principle. Consider any choice point within a decision tree, and the truncated tree that would then confront the person, were they to reach that point, i.e., the set of possible *subplans* that could be executed from that point on, together with their associated outcomes. Now construct a decision problem that is isomorphic to this truncated tree, i.e., presents the person with exactly the same set of possible sequences of choices, and associated outcomes, except that these choices are to be faced *de novo*, instead of taking place against the background of the prior choices that would have brought the person to that point in the original tree. Separability can then be formulated in the following manner:

> *Separability:* The subplan an agent would adopt at a designated point within the original tree (were she to reach that point), and subsequently execute, must correspond to the plan she would adopt, and subsequently execute, in the *de novo* version of the truncated tree that begins at the designated point.

Consequentialism together with separability, then, yields the standard result. The principle of consequences applied directly to a *de novo* decision problem involving a choice between just o_1 and o_2 at t_1 implies that o_2 will be chosen; and this, together with the separability principle, implies that at the second choice point in the problem given in Figure 1 she will choose o_2. Finally, it is the conjunction of these two principles

together with the principle of consistent planning that requires the rejection of plan $a_1 - a_3$.[7]

The self that is committed to the separability principle is an *incremental* reasoner: it is disposed to reconsider, at each new point in time, whatever plan was originally adopted, and to frame new plans on the basis of whatever, *then and there*, it judges will yield maximum expected-utility. Being an incremental reasoner is consistent with taking into account long-range considerations. On the model in question, the self at each point in time is presumed capable of projecting what, from its present standpoint, would best promote its long-term interests from that point on. Notice also that this sort of disposition to incremental reconsideration is not to be confused with that which is postulated in the standard model of revision in the light of new information: the self that reasons in accordance with the separability principle is disposed to change plans *even in cases in which there has been no change in relevant information*.[8] Separability also implies that the evaluation of any proposed coordination plan can appropriately proceed from the evaluation of the last segment of that plan, successively backwards, to the evaluation of the whole plan. Thus, consequentialism coupled with separability implies the classic "folding backward" principle of evaluation for decision trees.[9]

Separability places substantial restrictions on the capacity of an agent to cooperate with her own future self. Indeed, separability precludes intrapersonal co-ordination over time in any meaningful sense of that term. What is left to the agent who is committed to the separability principle is not coordination with her future self, but *strategic adjustment*. Her task is to determine how her own future self would choose, by reference to its own schedule of preferences, and then, to unilaterally adjust her own choice of plans, so as to maximize her preference or utility, given the specified constraints set by her own future self.

7 A much more detailed account of all that is involved in this proposed factoring of the conditions on rational dynamic choice can be found in McClennen (1990a).

8 For an illuminating discussion of planning that takes account of both the costs of. and the need for, reconsideration in the light of changing information, see Bratman (1987). See also, however, Bratman (1992) and the response by De Helian and McClennen (1992) for a discussion of some issues that are relevant to the thesis of the present paper.

9 To proceed in this fashion is to implicitly invoke what is known in the literature of dynamic programming as Bellman's Principle. See Bellman (1954).

II. The Paradoxical Nature of these Results

It is well established that a myopic approach to preference shifts can make the agent liable to what are clearly, from a consequential perspective, unacceptable outcomes. The extensive literature on Dutch-books and money-pumps shows that myopic choosers can be "tricked" into accepting bets and making other choices that result in a sure net loss of scarce resources. And since the myopic chooser's loss is the exploiter's sure gain, myopic choosers can fully expect that others will be eager to interact with them. All of this makes for a powerful argument against being myopic.[10]

It is considerably less appreciated that the sophisticated chooser faces a parallel set of liabilities. Yaari (1977) offers an illuminating diagnosis of this. He points out that the *intra*personal problem of dynamic consistency is formally equivalent, in certain important respects, to the paradigmatic problem of a non-zero-sum, non-cooperative game, i.e., to a certain class of *inter*personal choice problems. To suppose that the agent's preferences change over time is to suppose that the earlier self must contend with a later self that has a different schedule of interests. But that is tantamount to supposing two distinct selves that must interact with one another. The intrapersonal problem, then, can be modelled in terms of the standard interpersonal model. Now, on the standard account of the non-cooperative, *inter*personal problem, the rational solution will typically be an outcome that is *inter*personally sub-optimal (by the criterion of Pareto). By analogy, then, the *intra*personal dilemma can be understood as one in which the outcome of "rational" interaction with one's own future selves is intrapersonally *suboptimal*, i.e., each time-defined self does less well than it would have done, if the selves had really been able to effectively coordinate with each other. That is, sophisticated choosers typically have to settle for second best.

Consider once again the problem given in *Figure 1*. If plan a_2 is interpreted as a precommitment strategy, in which one hires an agent for a small fee, the inefficiency problem becomes particularly evident. On the assumption that each time-defined self is interested in having more rather than less financial resources, paying an agent means that

10 The extent and limits of these kinds of argument are discussed at great length in McClennen (1990a) and in McClennen and Found (1994).

each time-defined self is less well off than it would have been, if only it could have managed to serve as its own agent! Moreover, precommitment devices tend to be costly in two other respects. They limit the freedom of those who employ them, since they involve persons placing themselves in situations in which they do not act, and do not choose, but are acted upon, and have choices made for them. Moreover, they typically expose the agent to the risk associated with any procedure that is inflexible: agents can end up being unable to untie their hands in circumstances (projected or perhaps even not fully anticipated) where this can prove most disadvantageous. Ulysses, then, reduces his freedom by having himself tied to the mast, and also risks having his crew mutiny against him, while he is in that condition.

To be sure, that there are costs associated with being sophisticated is not necessarily a dispositive objection to that approach. If such costs are unavoidable—if it is not open to the agent to adopt any alternative approach to choice that is less costly—then the argument from costs goes nowhere. The key question, then, is whether the costs associated with a sophisticated approach are avoidable, whether there is some less costly alternative approach to sequential choice, and one that is open to an agent to adopt.

III. An Alternative Approach to Consistent Planning?

One can begin by recalling that the principle of consistent planning requires the agent to tailor her plans to a projection of what she will do in the future. This ensures consistency between present choice of a plan and future execution of a plan—consistency of a sort that makes the agent no longer liable to dutch-books and money pumps. In principle, however, consistency can be achieved in a radically different way.[11] Rather than regimenting present choice of a plan to projected future choice, the agent could regiment future choice to the originally adopted plan. Let us call an agent who manages to achieve consistency in this way: *resolute*. So characterized, being resolute involves being

11 An early version of this argument is to be found in McClennen (1985). Quite independently, Johnsen and Donaldson (1985) recognize the conceptual possibility of such an alternative way of achieving dynamic consistency. Having done this, however, they then proceed to focus their attention on what they take to be the more defensible strategy of sophisticated choice.

committed to carry out the plan that is selected. Thus, with regard to the problem given in *Figure 1*, a resolute chooser would be capable of choosing and then proceeding to execute, plan $a_1 - a_3$. Being resolute does *not* mean being unconditionally committed to never deviate from a chosen plan. Being faithful to a chosen plan need not take precedence in situations in which one acquires new information about outcomes. Being resolute only means that *if* one adopts a given plan (on the basis of one's judgment of its projected outcome), and *if* unfolding events are as one expected them to be, then one continues to adhere to that plan.

Now, resoluteness might, of course, express nothing more than a capacity of the agent to *tyrannize* over her own later selves. But resoluteness need not take this form: it can express instead the capacity of the agent to engage in sustained inter-temporal cooperation—to effectively coordinate with her own future and past selves, and this from an informed sense of what manner of choosing is required to avoid the problem of intrapersonal suboptimality. In all that is to follow, it is just and only this type of resoluteness that will be the focus of attention. Such a resolute chooser is consequentially oriented. In settling upon a plan of action, she compares the consequences of the various available plans, and rejects all plans that fail the test of intrapersonal optimality, as characterized by Yaari (1977).

Consider once again the version of the problem given in *Figure 1* in which plan a_2 constitutes a precommitment strategy of paying someone else to execute a choice of o_1 over o_2 at the second choice point. The reason for being resolute in this situation is, plainly and simply, consequential in nature: plan $a_1 - a_3$ is intrapersonally Pareto-efficient relative to plan a_2, the plan upon which the sophisticated chooser must settle. The suggestion, then, is that an agent who is faced with certain types of sequential problems will benefit from deliberating in terms of a two-level approach to consequentially oriented choice. Consequentially oriented considerations will guide her to adopt plans to deal with problems requiring intrapersonally coordinated choice, and these plans will then set constraints on subsequent choice. For such a chooser—a resolute chooser—the relevant question at any given choice point will be: is this particular option, as presented here and now, consistent with what is required with a view to an on-going (that is, previously adopted) plan, a plan that satisfies the criterion of intrapersonal optimality?

Sophisticated choosers must make arrangements for agents who will execute choices for them, or otherwise tie their hands in advance. In contrast, resolute individuals are able to *serve as their own agents* and dispense with precommitment devices. That is, those who can deliberate from such a more holistic perspective and then carry through with the plan judged best from that perspective, can achieve intrapersonal coordination without external enforcement mechanisms, precommitment props, and agency agreements. Such persons are able to thereby reduce the costs associated with executing the plan they most prefer.

IV. The Case For Resoluteness and Against Separability

It can be argued, of course, that the immediately preceding remarks speak merely to the implications of a conceptual possibility, and some of its implications, and that it remains an open question whether resolute choice can be defended within the framework of a theory of rational choice. The obvious problem is that being resolute implies violations of the separability principle. To appeal once again to the problem given in *Figure 1*, the agent who adopts and then resolutely executes the plan $a_1 - a_3$, despite being disposed to rank a_4 over a_3 in the context of *de novo* choice between a_3 and a_4 violates the separability principle. Since many are convinced that conformity to the separability principle is a necessary condition of rational choice, they are prepared to conclude that the model of resolute choice must be rejected. It recommends plans that are simply not feasible.[12] On the other hand, any argument for being resolute in such a context is ipso facto an argument against accepting the principle of separability in that same context.

Now, the scope of the separability principle is clearly limited in one respect. It is always possible that a particular agent just happens to intrinsically value making a commitment to plans. More generally, the description under which a given agent values some outcome may include reference to the manner in which this outcome is realized, so that from the value perspective of the agent the outcome cannot be characterized without reference to, and thus independently of, the path by which it is reached. The issue to be addressed, however, is whether path-independence (in this sense) must be rejected in cases other than

12 The problem of feasibility is discussed at length in McClennen (1990a: chap. 12).

simply those in which its denial is secured as the consequence of some *ad hoc* assumption about what the agent happens to prefer or value.

If one brackets these kinds of cases in which separability fails to hold, what can be said in favor of the principle? One argument turns on a consideration from the theory of action, to the effect that in so far as choice behavior is to be understood as determined by an agent's preferences (or valuations), it must be supposed that the preferences (or valuations) that function in this way are those that the agent has at the time of choice. Why? Because there can be, on the standard way of thinking about causal connections, no action at a distance. In itself, of course, this does not settle much, since it is always possible that the agent *now* has a preference for doing what, at some previous point in time, she intended to do. The "no action at a distance" argument, then, would become more significant if it could be shown that any counter-example to separability must appeal to some purely *ad hoc* assumption about present preference.

Conversely, to make the case against separability, one must show that a person can have a *reason*, as opposed to merely a preference, for taking into account past decisions, etc. This is a matter that will be addressed shortly. But before doing this, there is another quite distinct line of defense of separability to be considered. Some have been prepared to insist that the separability principle speaks to a fundamental requirement of *consistency*.[13] Descriptively, the root notion is that there should be a match or agreement between what one is prepared to choose at some particular point in a decision tree, and what one would choose if one were to face a parallel set of options *de novo*. The relevant question, then, is why this sort of match is required. Some are prepared at this point to simply appeal to "intuition." Unfortunately, what is clear from the past 40 years of debate is that many have just the opposite "intuition." But, surely, if this is all there is to be said, it is plausible to invoke a principle of *tolerance*, and let each theorist nurse his or her own intuitions. On this reading, however, no particular view can be accorded more than limited, inter-subjective standing, i.e., standing within the circle of the committed.

Perhaps it will be argued that it can be pragmatically or consequentially unfortunate for an individual to fail to accept the separability principle. That is, one might attempt to construct an argument parallel

13 See, for example, Hammond (1988).

to the ones that have been levelled against being myopic—dutch-book or money pump types of argument. This would serve to provide considerable leverage in favor of separability. Unfortunately, the thrust of the analysis so far is that it is not the *rejection* of separability, but the *acceptance* of separability that poses a pragmatic problem for the agent in a wide class of situations. That is, it is just and only the agent's own (putatively) fixed disposition to choose in a sophisticated manner that precludes her future self being able to coordinate with her former self and thereby implement what each self is prepared to acknowledge to be a more preferred outcome. Given this, to insist nonetheless that separability is a criterion of rational choice is to embrace the highly paradoxical conclusion that a fully rational agent, faced with making decisions over time will do *less well*—in terms of standard "economic" considerations of the conservation of scarce resources, freedom, and flexibility, than persons who are capable of a special sort of "irrationality"!

Given, however, that what stands between a sophisticated chooser and a more preferred outcome is just that agent's *own disposition* to choose in a sophisticated manner, the agent has a reason to alter her disposition—to reorient herself to a different way of approaching sequential choice problems. Moreover, and most importantly, this provides a way of rebutting the separability principle without having to appeal to a purely *ad hoc* assumption about what an agent just happens to prefer or value. The story just told is that a rational agent who fully grasps the logic of sequential decision problems can be led, by a sense of the gains to be secured—a consequential consideration—to settle on plans and take them as regulative for subsequent choice. In a behavioral sense, this means that they will choose, *as if* they valued carrying through on plans. But in this story there is no appeal to an *ad hoc* assumption. To the contrary, the suggestion is that being resolute can be grounded in a desire to effectively further one's own projects.[14]

Perhaps it will be claimed that the disposition to choose from a separable perspective is not easily abandoned. Consider, however, what decision theorists themselves have been prepared to say about

14 The argument here parallels one to be found in Gauthier (1986), in favor of what he characterizes as constrained maximization. McClennen (1988) offers some reasons for favoring the account given here over Gauthier's account, but relative to the objectives of the present paper, the differences between the two accounts are not very significant.

myopic choice. The claim is that there is a dispositive argument against being myopic and in favor of being sophisticated, an argument that is unabashedly consequentialist in its thrust: sophisticated choosers do better than myopic choosers. Moreover, the presupposition is that to become aware of the costs associated with being myopic, and to understand that these costs can be avoided by cultivating a sophisticated approach, is all that needed to convince one to be a sophisticated chooser. The case under consideration here, for being resolute rather than sophisticated, for deliberating from a non-separable rather than a separable perspective, is strictly parallel. Once again the argument is purely consequentialist in form. Resolute choosers do better than sophisticated choosers, in terms of husbanding scarce resources, in terms of flexibility, and in terms of freedom.

The argument just rehearsed does not establish that it is always rational to be resolute; it only establishes that resoluteness can be defended within the context of certain types of sequential decision problems. In particular, there is one setting within which the case for being resolute must be qualified. If the agent has doubts about her own capacity for making rational choices, this undercuts the case for being resolute. Moreover, nothing has been said about what, within a non-separable framework, a full theory of rational intrapersonal sequential choice would look like. All that has been argued so far is that there are contexts within which it is irrational, from a consequential perspective, to accept the separability principle.

This suffices, however, to yield a number of important conclusions. Consider once again Yaari's suggested analogy between intrapersonal and interpersonal choice. To advocate choosing in a resolute fashion is to reject Yaari's claim that the individual agent, faced with preference shifts over time, must settle for a second-best or intrapersonally suboptimal arrangement. That is, to reject the separability principle, and argue for being resolute, is to pave the way for the conclusion that *a self-consciously rational person will not fail to exploit the gains that can be realized through intrapersonal coordination.* This, in turn, has two connected, and quite powerful, implications. First, weakness of will is really a sign of imperfect rationality; and, secondly, talk about the principle of consistent planning and precommitment, and the like, is really best understood as addressed to those who are not, then, *fully* rational.[15]

15 The point here mirrors one that is to be found, although not always as explicitly as one would like, in Elster (1979). See in this regard McClennen (1990a: chap. 13.7).

V. Interpersonal Choice Under Ideal Conditions

What light does the foregoing analysis shed on problems of *inter*personal choice? Consider, first, the logically special case of games that are played under the following "ideal" conditions:

(1) all of the players are fully *rational;* and

(2) there is *common knowledge* of (a) the rationality of the players, (b) the strategy structure of the game for each player, and (c) the preferences that each has with respect to outcomes.

Just what is implied by (1) has yet to be spelled out, of course. With regard to (2) the assumption is not only that there is no asymmetry in the information available to the different players, but also that any deliberative conclusion reached by one player, regarding what choice to make, can be anticipated by the others as well: there are no hidden reasons.[16]

Here is a simple game of this type, one that involves the players choosing in sequence, rather than simultaneously: The first four outcomes are those that can be achieved by A and B coordinating on this or that plan (path through the tree). By contrast, the last outcome,

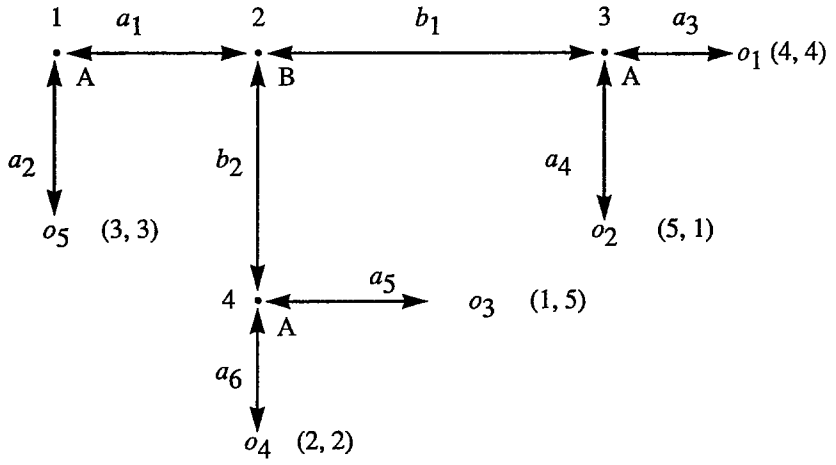

Figure 2: An Assurance Game

16 See Luce and Raiffa (1957: chap. 4.1) and Von Neumann and Morgenstern (1944: chap. 17.3).

associated with a_2, can be understood as one that is reached as the result of a unilateral move by A. Subsequently, an alternative interpretation will be considered: a_2 is a "precommitment" strategy, according to which B can be assured that if she chooses b_1 in response to a_1, a_3 will then be chosen, for an outcome of o_1, although at the cost of an "agency fee" to be paid to a third party from funds, say, contributed by A.

Given the specified preference rankings for outcomes, and the standard consequentialist assumption that plans are to be ranked according to the ranking of their associated outcomes, plan $a_1 - b_1 - a_3$ might seem to be the most likely candidate for a coordination scheme. To be sure, A prefers the outcome associated with plan $a_1 - b_1 - a_4$, but it is unrealistic to suppose that B would agree to coordinate on that plan. Under ideal conditions (of mutual rationality and common knowledge), however, talk of voluntarily coordinating their choices would seem to be pointless. If A were to set out to implement the first stage of such a coordination scheme, by selecting a_1, and if, for whatever reason, B were to reciprocate with b_1, what conclusion can then be reached about what A would select at choice point 3? There, so the argument goes, A would surely select a_4. Plan $a_1 - b_1 - a_3$ is simply not feasible: it calls upon A to make a choice that A knows she would not make, and, under ideal conditions, B knows this as well. Moreover, B ends up with her least preferred outcome, as the result of this attempt at "coordination." Suppose, then, that A were to select a_1, and B were to respond—on the grounds outlined above—by protectively selecting b_2: under these conditions, A's best response at choice point 4 would be a_6, and each would then end up with a second least preferred outcome. Once again, all of this is common knowledge. Against the background of these subjunctively characterized conclusions, then, A's best opening choice is not a_1, but a_2, yielding for each a third least preferred outcome.

Notice the way the reported reasoning proceeds here: from the last choices to be made in the decision tree, backwards, to the beginning of the tree. That is, under ideal conditions, each player can examine the last point at which some player has a choice to make; he can determine what, in terms of the stipulated preferences regarding the outcomes still available at that point, the player in question will choose there, and then incorporate that conclusion into a determination of what choice should be made by the player faced with making the immediately preceding choice. By working in this manner backwards through the

tree, each player can decide what to do at the first point at which she is called upon to make a choice.[17]

Now, the game given in *Figure 2* mirrors the essential features of the intrapersonal problem given in *Figure 1*. The outcome associated with $a_1-b_1-a_3$ is preferred by each to the outcome associated with a_2. But, according to the story just told, the former outcome is not *accessible*. Why? Not because of any disposition of B's, but because, as the story has been told, A would subsequently choose to depart from this plan. If B were sure that, upon arriving at choice point 3, A would select a_3 rather than a_4, B would be willing to choose b_1. B, then, would be interested in cooperating, were it not for the projected disposition of A to choose a_4 rather than a_3 at the later choice point. Thus, *A's quarrel is with herself*; or perhaps one should say, given the analysis of the last section, *with her own future self*!

What this suggests, of course, is that the argument of the last section can be applied to this situation as well. As already indicated, a consequentialist principle can be invoked to the effect that preferences for outcomes are controlling. And once again, it can be noted that this does not, in itself, settle the question of what would qualify as rational at choice point 3. What is requisite, in addition, is an assumption to the effect that it is only preferences for outcomes *still available at a given point* that are relevant to choice at that point. That assumption, in turn, can be secured by appeal to the very same *separability* principle introduced in the analysis of intrapersonal choice. It is consequentialism together with separability that once again yields the standard result. The principle of consequences applied directly to a *de novo* decision problem involving a choice between just the outcomes associated with a_3 and a_4 implies that the outcome associated with a_4 will be chosen; and this, together with the separability principle implies that at choice point 3 in the problem given in *Figure 2*, she will select a_4, and thus that, under ideal conditions, there can be no coordination on plan $a_1-b_1-a_3$.

It is separability that drives the form that backward reasoning, or "folding backward," takes in the analysis of sequential choice games. Separability implies that in evaluating any coordination plan, what that

17 The point here is that Bellman's Principle can be adjusted to apply also to sequential interpersonal choice problems.

plan calls upon a given agent to choose, at any given point, must be consistent with what that agent would choose, were she to make a *de novo* choice at that point. This is what licenses proceeding from the evaluation of the last segment of that plan, *taken in isolation from the rest of the plan*, successively backwards, to the evaluation of the whole plan. This is also what drives Selten's (1975) requirement that plans satisfy the subgame perfect equilibrium condition.[18]

The conclusion of the last section was that separability in the context of intrapersonal choice must be rejected. That conclusion carries over to the present context, where it is assumed that the game is played under ideal conditions. Once again, within the context of such sequential games, the appeal to separability with respect to one's own future choices cannot be defended. In the above game, no less than in the intrapersonal decision problems analyzed in the previous section, a commitment to separability precludes the agent from realizing gains that she could otherwise realize. Within this framework, then, separability cannot be taken as a criterion of rational choice.

The conclusion just reached is subject to a series of important extensions. *First*, It can be extended to certain simultaneous choice games played under ideal conditions. Consider, in particular, the following non-sequential version of game G, in which each player selects, *ex ante*, a plan of action for the sequential game, and the outcome of the game is then determined accordingly:

Player B→ Player A ↓	b_1	b_2
a_1/a_3, a_6	4, 4	2, 2
a_1/a_4, a_6	5, 1	2, 2
a_2	3, 3	3, 3

Figure 3: A Simultaneous-Choice Version of the Assurance Game

Once again, both players would be better off with the outcome which can be reached cooperatively, by A selecting plan a_1/a_3, a_6, and B selecting b_1, than with the outcome reached by unilateral choice of a_2 on the part of A.

18 The literature on subgame perfect strategies is extensive; for an excellent survey, see Fudenberg and Tirole (1992: chap. 3).

Notice that here, once again, A cannot plead that B's disposition to non-cooperation, were B to anticipate that A will choose cooperatively, requires A to choose instead the security maximizing unilateral strategy of a_2. B's maximizing response to a choice of a cooperative strategy by A is still to cooperate. To be sure, A does face here an assurance problem; but it derives solely from the consideration that B can be expected to choose defensively. Why? Because of A's dispositions. In this non-sequential game, then, just as in the single-person sequential decision problem discussed in the previous sections, A's problem is of her own making.

Indeed, under ideal conditions, it is impossible for B to rationalize a choice of b_2, except on the hypothesis that, as a rational player, A must choose plan a_1/a_4, a_6, given an expectation that B will choose b_1. That hypothesis is, of course, unavoidable within the framework of the standard account of rationality. What does the work here is not simply a consequentialist assumption to the effect that each player will be disposed to choose so as to maximize preferences for outcomes, but, once again, an assumption about how such consequential reasoning is to be *anchored*. The point is simply that however the process of deliberation has been conceptualized, it has been a fixed point of rational choice theory that a rational player cannot at the termination of deliberation make a choice which is not a preference maximizing response to the expectation she entertains with regard to the choice of the other player. What this means is that, whatever considerations frame deliberation, at the final moment of choice it is one and the same whether the other parameter affecting the outcome of a given player's choice is the choice of other player, or simply certain possible states of the world, about the occurrence of which she can form some more or less determinate estimate. In this way, the problem of interdependent choice is neatly reduced to a classic problem of *parametric* choice, i..e, of maximization of expected-utility against nature.[19]

19 There is a huge literature on refinements in, and modifications of, this way of thinking about rational interpersonal choice. What is basic is the concept of an equilibrium of choices, as developed originally by Nash (1951). The most useful exposition is still the one to be found in Luce and Raiffa (1957: chap. 4, 5). For some significant variations and modifications, however, see in particular Kadane and Larkey (1982). Bernheim (1986), and Harper (1991). It is also interesting to note, in this context, that Von Neumann and Morgenstern (1944) themselves partially abandon the parametric/equilibrium perspective when they move to the theory of *n*-person (as distinct from two-person) zero-sum games. In the case of a game between three or more players, there can be a parallelism of interests that makes cooperation desirable, and that will, in at least some cases, lead

This way of conceptualizing the problem of interpersonal choice implicitly invokes a separability requirement that is strictly parallel to the one invoked in the case of intrapersonal choice.[20] Stated somewhat more formally, the condition can be framed in the following manner:

> *Separability (for interpersonal, synchronous choice):* Let G be any game, and let D be the problem that a given player in G would face, were the outcomes of the strategies available to her in G conditioned, not by the choices of another player, but by some "natural" turn of events in the world; and suppose that her expectation with regard to those "impersonal" conditioning events corresponds to the expectation she has with regard to the choice that the other player will make in G: then her preference ordering over the options in G must correspond to the preference ordering she would have over the options in D.

Within the context of ideal games, this separability principle is, subject to precisely the same objection raised against the intrapersonal separability condition. Players who are disposed to choose in this fashion do less well, across a wide range of games, than those who are disposed to reason from a non-separable perspective, i.e., with a view to efficiently resolving the problem of sub-optimal outcomes. Here, then, is another context within which separable reasoning cannot qualify as rational—and where the rational solution to such games involves

to an agreement between some of the players involved. If the game is "zero-sum", of course, it cannot be in the interests of all the players to join in a grand coalition of all players, but smaller coalitions may still form. When this happens, von Neumann and Morgenstern imagine that the coalition will coordinate in such a manner as to secure the maximum payoff that it is possible for the members of that group to realize, thereby ensuring that between that coalition and those who remain outside, there will be a strict opposition of interest. There is, then, a real place for cooperation within their theory of n-person, zero-sum games. They also sketch a theory of non-strictly competitive games that retains the presupposition that rational agents will be disposed to coordinate when there are gains to be secured thereby. In particular, they suppose that any non-strictly competitive game involving n agents can be embedded in a strictly competitive game in which there is one additional "fictional" player—can this not be thought of as nature?—whose payoff is simply the negative of the payoff that the n players can achieve if they form a grand coalition. This, in turn, would imply that for the classic version of the prisoners' dilemma game, played under ideal conditions, the two agents can think of themselves as jointly playing a strictly competitive game against nature, where their best strategy is to fully cooperate with one another, and thereby force the maximum joint payoff possible from nature.

20 This formulation of separability as a condition on ideal simultaneous choice games is explored at greater length in McClennen (1992).

resolute cooperation on the part of a player who finds herself in a position parallel to that of A in the game just analyzed.

This immediately suggests a *second* extension, however, to the more familiar symmetrical prisoners's dilemma games. Consider, for example, the following game:

Player B→ Player A ↓	cooperate	defect
cooperate	4, 4	1, 5
defect	5, 1	2, 2

Figure 4: Prisoners' Dilemma

Each player is in a position to exploit a cooperative response by the other to personal advantage, by unilaterally defecting; but mutual defection yields less for each than would mutual cooperation. On the standard account, mutual defection is the only rational solution, (even) under ideal conditions. In such a game, however, a given player only faces an assurance problem in virtue of the conclusion that expected cooperation on her own part will trigger a decision to defect by the other player. That conclusion, in turn, pivots once again on the assumption that a rational player must be committed to the separability principle. Once that assumption is dropped, mutual cooperation is clearly the rational solution to the prisoners' dilemma game.

Consider, now, a *third* possible extension, to the case of n-person non-iterated games (where $n > 2$), played under ideal conditions. Here, it might seem, there is a complication that can serve to rationalize defection, at least for some. Take a classic "public-goods" problem in which all would do better if all cooperate, than if none cooperate, but some can do even better if they do not cooperate, so long as a sufficiently large number of others do cooperate. The suggestion, in effect, is that once a sufficiently large proportion of the players are clearly disposed to cooperate, such an "arrangement" is rationally exploitable by "non-cooperators."

How is this conclusion to be understood? Supposing that free-riders do better than those who continue to cooperate, how is one to generate, within such a model world, the conclusion that some act rationally by free-riding, while others act rationally by cooperating? By what marks

are the rational defectors to be distinguished from the rational cooperators? Such distinguishing marks will have to be found. It cannot be rational for each and every participant to free-ride, for it is a characterizing feature of such a model that for some it must be rational to cooperate.[21]

A *fourth* possible extension is to *n*-person *iterated* games. Suppose that one grants the rationality of cooperating, under ideal conditions, in an *n*-person simultaneous choice, single stage game. So long as the condition of mutual rationality and common knowledge is preserved, there is no bar to extending that result to *finitely* iterated versions of such a game. In particular, then, the conclusion reached above regarding the rationality of adopting a non-separable approach to interdependent choice implies, in turn, that the "sub-game" perfectness condition, and other artifacts of the standard backward induction argument, must be rejected.[22]

What about *infinitely* iterated games (or games whose terminal stage is determined probabilistically)? Interestingly enough, there are a variety of folk-theorems that leave open the rationality of fully "cooperative" interaction in such games—where "rationality" is interpreted in the standard fashion (in terms of some refinement or other of the equilibrium condition, or, more generally, the concept of choice that maximizes against rational expectations).[23] These theorems typically

21 One can imagine, of course, a model in which some irrationally cooperate, and others, then, rationally take advantage of them. But that does not qualify as a model in which all are making rational choices. In principle, of course, it may be possible to characterize a model of interaction under ideal conditions where specific, and asymmetrical, features of the situation permit the rationalization of differential behavior. Perhaps, for example, some are in a position to issue much more credible threats than others, and that this gets reflected in a differential apportionment of responsibilities, etc. But in so far as a potentially disadvantaged player can anticipate that such situations will be iterated, it is still an open question whether acceptance of a less advantageous distribution is rational. Here, once again, there may be a rational role for policies of resentment and resistance. See, for example, McClennen (1989, and 1990b), and also the interesting treatment of reason and passion in Frank (1988).

22 Even if some version of backward reasoning can be sustained, it will have no bite in the context of the games under consideration. That is, one will not be able to unravel the argument for cooperation in the iterated version, by appeal to the rationality of defecting on the last round—for on the account just given, it is rational to cooperate on the last round, and also on the second to last, etc.

23 Once again the literature here is extensive. For an excellent overview, see Fudenberg and Tirole (1992).

turn, however, on very complex and rather problematic assumptions about how a hierarchy of *credible* threats of punishment can be utilized to make conformity to the coordination scheme rational. In addition, the folk theorems do not smoothly translate into the conclusion that just and only Pareto-optimal outcomes can be rationalized. Indeed, what the theorems establish, rather distressingly, is that virtually *any outcome can be rationalized.* By way of contrast, to make the case for thinking non-separably about rational interaction in such games, is to establish directly the rationality of cooperating on a Pareto-optimal outcome. This, moreover, is a conclusion that can be reached without invoking any problematic assumptions about a hierarchy of credible threats.

More generally, the conclusion just reached regarding interaction under ideal conditions enables one to close what has heretofore constituted a significant gap between the theory of ideal, non-zero-sum games, and the microeconomic theory of interaction under ideal conditions of perfect competition. It is a central insight of the theory of perfectly competitive market exchange that rational agents who know each other to be such will not fail to exploit the mutual gains that trade makes possible. That is, they will act in a way that ensures that the condition of Pareto-optimality will be satisfied. On the theory proposed above, this will hold for other forms of interaction as well.

In the case of the model of perfect competition, of course, satisfaction of the optimality condition is the unintended outcome (a mere by-product) of bilateral exchange between interested sellers and buyers. In the case of *n*-person cooperative ventures, by way of contrast, what is presupposed is a tacit or explicit *n*-person agreement (or understanding) on the terms of the cooperative venture, an agreement the intended point of which is to ensure that mutual gains do not remain unexploited. This suggests that a theory of *n*-person rational interaction under ideal conditions cannot be completed without a theory of bargaining. That is, contrary to the direction taken by most game theorists in the past few decades, it is the theory of cooperative, rather than non-cooperative games that must now become the focus of attention.[24]

24 This is something that Gauthier (1986) has made admirably clear in his argument for constrained maximization. See also McClennen (1989 and 1990b).

VI. Interpersonal Choice Under More Realistic Conditions

Only a few, perhaps, will be prepared to accept the conclusions reached above regarding ideal games. But even those few are likely to reject, out of hand, any suggestion that these results can be extended to more realistic models of interpersonal choice. Why is this? What is it about "more" realistic cases that poses the problem?

One can begin by noting two relevant considerations. In the two-person games discussed above, the rational disposition to cooperate is clearly only a disposition to *conditionally* cooperate: each is willing to cooperate, *given* well-defined expectations regarding the cooperative dispositions of the other players. Under realistic conditions of interaction, however, the requisite expectations regarding the other players are not so easily secured. It may simply happen that among those who are interacting, some are not rational; alternatively, all may be rational, but this may not be common knowledge. Thus the representative participant will have an assurance problem. In the real world, the assurance problem arises in a wide variety of situations. Regardless of the rationality, in principle, of cooperating, individuals often confront situations in which uncertainty regarding the rationality of other participants, or of other features of the game, requires them to proceed cautiously. Under such conditions, cooperation is not easily secured.

There is a *second* respect in which the common knowledge assumption can fail and which also seems to pose an assurance problem. Under ideal conditions, it is transparently clear what choice each identifiable player makes, if only after the fact. In the real world, however, conditions of relative anonymity often prevail. That is, the form of interaction may simply preclude identifying who did what, or, at least doing this cost-effectively. In such cases, must one conclude that it is rational for anonymous choosers to defect?

To isolate the relevant consideration, suppose that all the participants are known to be rational, and that there is common knowledge of this, and of other aspects of the interactive situation. That is, suppose that common knowledge fails *only* in the respect that some one or more identifiable participants cannot be linked to their chosen acts. If such anonymity is symmetrically distributed among the participants, this case is not relevantly distinguishable from the ideal case: if all defect, all are worse off than if no one defects, and given that anonymity is symmetrical, there is no rational basis on which to recommend that some, but not others defect.

One can recall here 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. Change the story, however, so that each of a number of persons who interact with one another possesses such a ring. Suppose, in keeping with the type of interdependent choice situation under consideration, that if all use their rings, the result will be mutually disadvantageous. That is, suppose that whatever a given person gains by using the ring is more than offset by the losses incurred as result of others using their rings. Each, then, will stand to gain if all refrain from using their rings.[25] Now, this is, in fact, a fair description of the real world. Each agent can expect to find herself in situations in which her actions are relatively invisible to others. In this symmetrical setting, however, the conclusion reached above for the ideal case carries over. From a purely consequential perspective, the disposition to defect in such situations, cannot be characterized as a rational disposition.

Could it be argued that, in the case of *symmetrical* anonymity, each participant still faces an assurance problem? Not if anonymity is the only respect in which the common knowledge assumption has been relaxed! On the assumption that there is common knowledge of each other's rationality, and of the other relevant features of the game, either rationality prescribes that each defect, in which case there is no problem of "assurance" (mutual defection is a foregone conclusion) or it prescribes cooperation, in which case each has the assurance she needs!

What if anonymity is *asymmetrically* distributed? Asymmetry of this sort would place some, but not all, in a position to exploit others. Under those conditions, on the standard account, it is rational for those who are in a position to exploit others to do so. This conclusion, however, appears to hinge critically on the consideration, already explored within the context of ideal games, that, even in the face of rational defection on the part of some, for others the rational choice will be to cooperate (with one another). If the rational response by conditional cooperators

25 Of course, for any given player, using the ring dominates not using the ring, and this, on the usual account, yields the conclusion that it is rational for each to use her ring. The observant reader will note, then, that a theory of non-separable reasoning leads to the rejection of dominance conditions.

to anonymous defections is to join the ranks of the defectors, and this is common knowledge, the anonymous defectors precipitate a situation in which all, including themselves, are worse off than they could have been. But in this case, the disposition of "early" defectors cannot be rationalized.

Suppose, however, that the structure of the situation is such that continued cooperation by most is rational even in the face of defection by some. Even here, there are still threshold problems that may confront would-be defectors. If the number of those who are in a position to choose anonymously is greater than the number who can defect with impunity, it cannot be concluded smoothly that anonymity as such rationalizes defection. The hard case, then, is the one in which anonymity is asymmetrically distributed *and* the number of those who are in a position to defect anonymously is less than the threshold number that would trigger a rational decision on the part of the remaining conditional cooperators to defect. In such a case, it might seem that a defection strategy can be rationalized, and non-defectors will be exploited.

But even here, the rationality of defection cannot be smoothly defended. If those who are disposed to cooperate thereby expose themselves to being exploited, the rational response for them will be to expend additional resources on targeting exploiters, and, if it is necessary to ensure that such expenditures are made, even to cultivate a special hostility towards exploiters. Once again, the point is simply that while a disposition to root out free-riding exploiters, and retaliate against them, will generate choices that cannot be rationalized incrementally, certain dispositions of this sort will be rationalizable from a more holistic perspective. The cultivation of a "retaliatory" approach to dealing with would-be exploiters implies in turn that sophisticated exploiters will incur significantly greater risks—both in terms of the severity of the penalties they will suffer if exposed, and the likelihood that they will be exposed. This means, typically, that the range of situations in which an exploitative disposition is rational will be reduced. That is, parametric reasoners who are poised to exploit others will have to contend with a lower expectation and a greater variance in return. Even, then, if there are conditions under which defection strategies can be rationalized, it may still be possible to show that cooperators will have an interest in transforming those conditions so as to minimize the occasions on which exploitation can be rationalized.

Alternatively, cooperators can attempt to cluster together and exclude exploiters.[26]

Moreover, and most importantly, to acknowledge that in the "real" world circumstances may make it rational for some to defect, is *not* to acknowledge that it is rational for the representative participant to develop and refine a capacity to dissemble—to cultivate the appearance of being a cooperator, while in fact being fully disposed to defect whenever this can be done with impunity. One cannot conclude the latter, because it is addressed to the representative participant. That is, were it systematically be acted upon, the result would be consequentially unacceptable to each and every participant.[27]

VII. Rules, Resoluteness, and Rationality

It is time to return to the question of the rationality of accepting constraints on one's choice, where the rules defining those constraints constitute practices in Rawls' sense of that term. It was argued above that the model of separable reasoning provides neither a secure footing for the concept of, nor a rationale for, choosing subject to the constraints of practice rules. Does the model of consequentially based, non-separable, holistic deliberation fare any better?

Consider first the *concept* of a practice. One can mark in the abstract concept of being resolute a model for just the sort of hierarchical structuring of choice that characterizes practice-constrained choice. This parallel obtains, moreover, even if resoluteness is conceived as merely the imposition, by the earlier self, of a regimen upon the later self, or, in the case of interpersonal choice, of a convention, among a group of people, regarding how each will constrain her choice in certain situations—that is, where the question of the rationale of doing this sort of thing has been bracketed.

The abstract concept of being resolute also provides a model for a particular logical feature of a practice, namely, that the rules defining the practice must be understood as "prior" to the cases to which those rules apply. How is this notion of priority to be understood? Rawls himself explicates it by appeal to the status of the rules defining a

26 The point here parallels the one made in footnote 21 above.

27 This, hopefully, is the way in which one can disarm Sayre-McCord's (1989) otherwise powerful argument for strategies of deception.

game—in the ordinary sense of that term, e.g., the game of baseball. The idea is that the question of the appropriateness of constraining one's behavior in the manner directed by the rules of an applicable practice cannot even be raised except against the background supposition that such a practice exists. Indeed, deliberation that by-passed the rules of the practice, and appealed directly to the underlying considerations upon which the rules themselves were grounded, would make no sense at all.[28]

It is clear from Rawls' discussion as a whole, however, that the priority of the rules of a practice to cases falling under those rules has meaning even when the rules do not define a game, in the ordinary sense of that term, but simply mark out a form of activity in which choice is to be understood as rule governed, rather than purely discretionary, in character. The sense of priority that is appropriate in those cases turns on the notion that the very nature of the choice that confronts one is at least partially determined by whatever practices are *in effect*. Here, unlike the case of the rules of a game, it might very well be true that one could make perfectly good sense of what one's options were, and have a basis for evaluating these options, even if there were no extant rule governing the case. Yet it is the existence of a practice rule that frames the problem for one. That is, in the absence of the practice rule, one would understand oneself as facing a very different problem, a different set of considerations as pertinent to the evaluation of the options.

It is this sense of priority that is relevant to rule governed choice in the areas both of intrapersonal and interpersonal policy deliberation and decision. The sense in which the choice to be made in some sequential decision problem is understood as constrained by a prior decision to pursue a plan, or a prior (tacit or explicit) understanding as to how choices by quite different individuals are to be coordinated, is the sense in which a practice rule establishes non-discretionary constraints on choice. What is relevant to subsequent choice is not what plan one might have adopted, or what plan it would have been best to adopt (by reference to some underlying consideration), but what plan one did in fact adopt. And, correspondingly, what is relevant to certain classes of choice is not what practices one might have adopted, or what practice

28 See Rawls (1955: 190).

it might have been best to adopt (once again, by reference to some underlying consideration), but what practices are in effect.

The idea of non-separable deliberation provides, however, much more than simply a framework within which to make sense of the notion of intentionally choosing in accordance with the constraints of a practice. It also provides a framework within which a case can be made for having practices, for hierarchically structuring choice by reference to the rules of a practice. Once again, the logic of such rules involves the notion that one cannot decide to overrule such a constraint in a given situation to which the practice rule applies by directly appealing to whatever considerations support the practice itself: those who participate in such a practice abdicate the right to make decisions case by case by direct appeal to such underlying considerations. Nonetheless, it is central to Rawls' account that there are such underlying considerations, and that the nature of those considerations supports such a hierarchical structuring of choice.

Rawls' remarks in this regard are brief, but what he does say is adaptable to the present discussion. The suggestion is that there are cases in which the concerns of each cannot be served unless the future is tied down and plans coordinated in advance. The notion is that in such cases each person's deciding what to do by reference to her own concerns, case by case, will lead to confusion, and the attempt to coordinate behavior simply by each trying to predict the behavior of the others will fail. When such conditions obtain, and when persons are thus led by these concerns to agree to a coordination scheme, what they agree to is a practice.

On this sort of account, there is no need to introduce some *ad hoc* assumption about persons just happening to attach value to choosing in accordance with such rules. Nor are such persons "rule bound" in a way that can be criticized from the perspective of a theory of consequential choice.[29] The story to be told here pivots on consequential concerns. It is a story of individuals who come to regulate their interactions with themselves over time, and with one another, in accordance with constraints to which time-indexed selves, or distinct individuals, can mutually assent, and do this from nothing more than a sense

29 For a rebuttal to the charge that being resolute must involve being "rule bound," see De Helian and McClennen (1993. forthcoming).

of the enhanced power that such a new form of activity gives them with respect to furthering their own projects and interests.

IIX. The Issue of Constitutional Choice

Consider now what the various arguments presented above imply with regard to constitutional arrangements. In very general terms, the model of constitutional choice that is central to the work of Buchanan and other public choice theorists reveals itself, on close examination, to embody just the kind of hierarchical structuring of choice by reference to practices with which the present analysis has been preoccupied. The foregoing analysis, then, points the way to how, in principle at least, an "ethics of rules" which is an essential component of an *effective* social or constitutional charter can be defended; and, most significantly, defended without having to make an appeal to exogenously specified moral norms or the like. The crucial move involves modifying the conception of rational, consequentially oriented choice so as to provide room for a more holistically oriented consequentialist perspective. At root, moreover, the argument pivots on a simple observation: what holistic reasoners are able to achieve constitutionally, by virtue of being able to participate in rule governed practices, is all that incremental, parametric reasoners can achieve, but only more effectively, as measured in terms of those primary goods with which decision-theorists and economists have always been preoccupied: material resources, flexibility and liberty.

Take, for example, the issue of constitutional restrictions on legislative and executive action. The argument is that such restrictions derive their validity or authority primarily from the consideration that individuals who enter into political interaction with one another need assurances that certain of their concerns will not be overridden by political coalitions that have sufficient power (in terms, say, of majoritarian voting rules) to prevail. But these mutual concerns support a hierarchical system of choice: what is recommended is not merely a set of cautionary maxims for, but a set of formal side-constraints on, political and economic choice. This impies, in turn, that there can be no direct appeal to mutual concerns for the purpose of setting aside constitutional restrictions in specific cases, although it is still open to participants to amend or otherwise alter their constitutional arrangements. Again, however, while individuals can debate with one another whether some

modification in these restrictions might be mutually acceptable—it is also an implication of the above analysis that individuals must be prepared to discipline themselves to refrain from rent seeking attempts to modify these constraints in ways that work to their own advantage but to the disadvantage of others.[30]

But, of course, the need for hierarchically structured choice procedures extends well beyond the domain of sheltered rights (as expressed, say, in the original Amendments to the United States Constitution), for it is of the essence of any legal and/or political system that it systematically structures the interactions of its participants in terms of practices. Buchanan's *The Limits of Liberty* (1975) provides a sustained diagnosis of the many pitfalls into which individuals fall when their pursuit of interest happens to be undisciplined by an "ethics of rules." In the first place, an individual will often have a rational incentive to violate the rules defining these arrangements. One can try to correct for this by threatening violators with sanctions, but this simply generates a new set of problems. Enforcement mechanisms require the expenditure of scare resources, and involve a loss of liberty and privacy. Moreover, in so far as the point of employing sanctions is to secure compliance, administrators may well be reluctant to apply them, when the threat of their use fails to effect compliance. But, of course, potential defectors will anticipate such reluctance. This, in turn, will lead to either increased violations, or increased levels of surveillance and more severe sanctions. In addition, resources will also have to be expended to conserve the various constitutional institutions themselves, since rational individuals will have an incentive to not simply violate the terms of these arrangements, but to alter them in ways that work to their own advantage but to the disadvantage of others. Finally, given the disposition of rational individuals to both free-ride and rent-seek, voluntary (market) contracts will have to be supplemented with collective choice mechanisms. But rational *political* individuals will end up voting for an oversupply of public goods. And, once again, the attempt to resolve this problem by means of constitutional constraints is likely to be frustrated by the rent-seeking activities of rival interest groups.

As certain of these remarks suggest, the revised model of rational, *intra*personal choice, as presented in Sections I. through IV., has direct and independent application to constitutional choice. First, with regard

30 The case for this is discussed at length in (1989, and 1990b).

to long-range political and economic policy, it provides a powerful way of underpinning, and extending, the argument presented in the path-breaking work of Kydland and Prescott (1977), regarding the importance of binding rules rather than discretionary approaches for the implementation of sound long-range political and economics policies.[31] The argument for being resolute takes their line of reasoning to its logical conclusion: If it serves the interests of the members of a society to have its officials precommit to certain policies, and thereby avoid the costs associated with incrementally exercised discretion, then it will serve those interests even more to have such officials resolutely implement these policies. Second, and more generally, what is required for effective implementation of any constitutional arrangement between rational cooperators is not simply the capacity to show restraint with respect to ever-present opportunities for individuals and groups to rent-seek, but also the capacity to resolutely resist the rent-seeking efforts of others. Finally, a theory of resolute choice provides a grounding for a policy of executing threats. The point is that with regard to any constitutional or post-constitutional practice, a known capacity to execute threats, even when they have failed to accomplish their purpose, serves to increase the chances that such threats will accomplish their purpose.[32]

It is clear, of course, that in order to make the case for an "ethics of rules" that has implications for constitutional political economy, one must move beyond simplified models of interaction under ideal conditions. The relevant model is presumably one that involves complex interaction between a very large number of persons, under "realistic" conditions of uncertainty or limited knowledge about crucially

31 Kydland and Prescott (1977) has emerged as something of a classic for the treatment of this and related issues of rules v. discretion. See, for example, *The Economist* 2 March 1991: 71–72, for a survey; and also *Federal Reserve Bank of Philadelphia Business Review* March 1985.

32 Buchanan (1975a) contains an important analysis of this type of problem. On the occasion of the original presentation of Buchanan's paper, the present author chose to demur somewhat from Buchanan's conclusions (see McClennen 1977). The present paper, then, gives expression to a belated, but nonetheless required, retreat from an earlier view, and a recognition of the central importance of the point that Buchanan sought to make. This author is still persuaded, however, that some of the particular examples that Buchanan chose to focus upon—that of students v. college administrators, and welfare recipients v. government officials—pose special problems of justice that remain to be dealt with, notwithstanding the force of Buchanan's argument.

important variables, including whether those others with whom a given agent interacts are conditional cooperators, and the presence of asymmetries of a kind that may even bring into question the rationality of being a conditional cooperator. Given all this, one might be tempted to infer that the foregoing analysis (at best!) only implies something about how very small groups of ideally rational persons might, under highly unrealistic conditions, interact with one another.

It can be argued in reply, however, that the models particularly relevant to constitutional and post-constitutional systems of rules do not involve random, essentially non-repeatable bilateral encounters, where the problem of assurance is most severe; nor are they those in which effective targeting of defectors proves too costly. On the contrary, for the purpose of analyzing the problem of efficiently sustaining constitutional structures, the most relevant model is that of a dynamic *n*-person game with randomly determined terminal states. Within the typical constitutional setting, defections from established practices are visible, choices of action are sequential, and problems of "endgame" play do not arise. In that setting, the most pressing of the problems discussed above—the assurance problem—is surely one that can be resolved. In particular, one can optimistically project here, at the level of constitutional choice, a progressive development. At a first stage, it can be supposed that individuals will employ various traditional enforcement devices to overcome the assurance problem, and thereby manage to generate a history of cooperation. Subsequently, it can be supposed that such individuals will then be led, by a sense of the mutual gains to be realized, to let such costly props gradually fall away and be replaced by "principled," that is to say, rule governed, choice.

A Closing Thought

The conclusion just reached is that the standard way of thinking about both intrapersonal and interpersonal choice is defective, and that there is a better theory waiting to be developed around the concept of holistic or non-separable ways of evaluating plans and coordination schemes. This is, moreover, a theory that can not only make sense of the notion of a practice, but provide a consequentialist rationale for a commitment to practices. Beyond theory, however, it is worth pondering on what might be the effect of a course of study in which, say, at the very minimum the issue of what rationality requires in such choice

situations was not begged in favor of the extremely limiting sort of model to which one is driven by the separability principle, and the parametric form of reasoning it implies. Suppose, in particular, that a more concerted effort were made to make individuals aware of the complex nature of decision-making over time, and in interactive situations with other agents, and to at least mark out concretely the advantages to be realized by those who could resolutely serve as their own agents, and choose within the context of cooperative schemes in a principled fashion. One might then expect to see this more efficient mode of decision-making drive out more costly precommitment and enforcement methods, and this through nothing more than what economists like to describe as *the ordinary competitive process*.

REFERENCES

Ainslie, G. (1993) *Picoeconomics*. Cambridge, Mass.: Cambridge University Press.

Bellman, R. (1954) "The Theory of Dynamic programming." *Bulletin of the American Mathematical Society*. 503–515.

Bernheim, B. D. (1986) "Axiomatic Characterizations of Rational Choice in Strategic Environments." *Scandinavian Journal of Economics* 88: 473–488.

Bratman, M. E. (1987) *Intention, Plans and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Bratman, M. (1991) "Planning and the Stability of Intention." *Minds and Machines* 1.

Brennan, G., and Buchanan, J. (1985) *The Reason of Rules*. Cambridge, Mass.: Cambridge University Press.

Buchanan, J. (1975a) "The Samaritan's Dilemma." In: Phelps, E. S. (ed) *Altruism, Morality, and Economic Theory*. New York: Russell Sage Foundation.

Buchanan, J. (1975b) *The Limits of Liberty*. Chicago: The University of Chicago Press.

Buchanan, J. (1979) "Natural and Artifactual Man." In: Buchanan, J. *What Should Economists Do?* Indianapolis, Indiana: Liberty Press.

Buchanan, J. (1991a) "Economic Interdependence and the Work Ethic." In: Buchanan, J. *The Economics and the Ethics of Constitutional Order*. Ann Arbor, Michigan: The University of Michigan Press.

Buchanan, J. (1991b) "Economic Origins of Ethical Constraints." In: Buchanan, J. *The Economics and the Ethics of Constitutional Order*. Ann Arbor, Michigan: The University of Michigan Press.

De Helian, L., and McClennen, E. F. (1992, forthcoming). "Planning and the Stability of Intention: A Comment." *Minds and Machines* 2.

Elster, J. (1979) *Ulysses and the Sirens*. Cambridge, Mass.: Cambridge University Press.

Frank, R. (1988) *Passions within Reason*. New York: W. W. Norton & Co.

Fudenberg, D., and Tirole, J. (1992) *Game Theory*. Cambridge, Mass.: The MIT Press.

Gauthier, D. (1986) *Morals by Agreement*. Oxford: Clarendon Press.

Hammond, P. (1976) "Changing Tastes and Coherent Dynamic Choice." *Review of Economic Studies* 43: 159–173.

Hammond, P. (1977) "Dynamic Restrictions on Metastatic Choice." *Economica* 44: 337–350.

Hammond, P. (1988) "Consequentialist Foundations for Expected Utility." *Theory and Decision* 25: 25–78.

Harper, W. (1991) "Ratifiability and Refinements (in Two-Person Noncooperative Games)." In: Bacharach, M., and Hurley, S. (eds) *Foundations of Decision Theory*. Oxford: Basil Blackwell.

Hobbes, T. ([1651] Library of Liberal Arts Edition 1958) *Leviathan*. Indianapolis, Indiana: The Bobbs-Merrill Company.

Hume, D. ([1739/40] Selby-Bigge Edition 1888) *A Treatise of Human Nature*. Oxford: Oxford University Press.

Johnsen, T. H., and Donaldson, J. B. (1985) "The Structure of Intertemporal Preferences Under Uncertainty and Time Consistent Plans." *Econometrica* 53: 1451–1458.

Kadane, J. B., and Larkey, P. D. (1982) "Subjective Probability and the Theory of Games." *Management Science* 28: 113–120.

Kydland, F. E., and Prescott, E. C. (1977) "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85: 473–491.

Luce, R. D., and Raiffa, H. (1957) *Games and Decisions*. New York: Wiley.

McClennen, E. F. (1977) "The Samaritan's Dilemma: Comment." In: Phelps, E. S. (ed) *Altruism, Morality, and Economic Theory*. New York: Russell Sage Foundation.

McClennen, E. F. (1988) "Constrained Maximization and Resolute Choice." *Social Philosophy & Policy* 5: 94–118.

McClennen, E. F. (1989) "Justice and the Problem of Stability." *Philosophy & Public Affairs* 18: 3–30.

McClennen, E. F. (1990a) *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge, Mass.: Cambridge University Press.

McClennen, E. F. (1990b) "Foundational Explorations for a Normative Theory of Political Economy." *Constitutional Political Economy* 1: 67–99.

McClennen, E. F. (1992) "The Theory of Rationality for Ideal Games." *Philosophical Studies* 65: 193–215.

McClennen, E. F., and Found, P. (1994, forthcoming). "Dutch-Books and Money Pumps." *Theory and Decision*.

Nash, J. F. (1951) "Non-cooperative Games." *Annals of Mathematics* 54: 286–95.

Neumann, J. v., and Morgenstern, O. ([1944] 1953) *Theory of Games and Economic Behavior*. New York: Wiley.

Pollack, R. S., (1968) "Consistent Planning." *Review of Economic Studies* 35: 201–208.

Peleg, B., and Yaari, M. E. (1973) "On the Existence of a Consistent Course of Action When Tastes Are Changing." *Review of Economic Studies* 40: 391–401.

Rawls, J. (1955) "Two Concepts of Rules." *Philosophical Review* 64: 3–32.

Sayre-McCord (1989) "Deception and Reasons to Be Moral." *American Philosophical Quarterly* 26: 113–122.

Schelling, T. (1978) "Egonomics, or the Art of Self-Management." *American Economic Review: Papers and Proceedings* 68: 290–294.

Schelling, T. (1984) *Choice and Consequence*. Cambridge, Mass.: Harvard University Press.

Strotz., R. H. (1956) "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies* 43: 149–58.

Vanberg, V., and Buchanan, J. (1990) "Rational Choice and Moral Order." In: Nichols, J. H., Jr., and Wright, C. (ed) *From Political Economy to Economics and Back?* San Francisco: ICS Press.

Yaari, M. E. (1977) "Endogenous Changes in Tastes: A Philosophical Discussion." *Erkenntnis* 11: 157–196.