Reasonable Utility Functions and Playing the Cooperative Way

Gerald F. Gaus

1. The Rational, the Reasonable and Utility Maximization

In *Political Liberalism* Rawls draws the important distinction between the rational and the reasonable. The rational

> applies to a single, unified agent (either an individual or corporate person) with the powers of judgment and deliberation in seeking ends and interests peculiarly its own. The rational applies to how these ends and interests are adopted and affirmed, as well as to how they are given priority. It also applies to the choice of means, in which case it is guided by such familiar principles as: to adopt the most effective means to ends, or to select the more probable alternative, other things equal….
>
> What rational agents lack is the particular form of moral sensibility that underlies *the desire to engage in fair cooperation as such*, and to do so on terms that others as equals might reasonably be expected to endorse. I do not assume the reasonable is the whole of moral sensibility; but it includes the part that connects with the idea of fair social cooperation (Rawls, 1996: 50-51, emphasis added).

Because the rational and the reasonable are distinct in these ways, says Rawls, it is a mistake to try to derive the moral (qua the reasonable) from the rational; thus he appears to criticize David Gauthier's project of basing morality on the theory of rational choice qua utility maximization (even though it was Rawls's earlier work that inspired Gauthier; Rawls, 1996: 53; Gauthier, 1986: 4). Others have followed up this idea, and have argued that, given decision theory's focus on instrumental rationality qua the maximization of utility, decision theory cannot adequately capture the ideas of the moral or the reasonable insofar as they manifest "the desire to engage in fair cooperation as such." In this vein, Paul Clements and Emily Hauptmann (2002) argue that, while decision theory's modeling of the rational leads to problems such as the Prisoner's Dilemma (PD), drawing on the reasonable allows us to make sense of playing a PD in a cooperative way.

In this essay I dispute this conception of the utility and decision theory, which ties it to means-end, instrumental, reasoning. I show that the decision theoretic framework has no deep problems accommodating the "reasonable" qua a desire to engage in fair cooperation as such. I focus on the claim that, while rational choice-driven agents are caught in the Pareto-inferior outcome, reasonable agents could "solve" the PD and cooperate. Not so, I shall argue. All evaluative criteria relevant to choice can be built into a von Neumann-Morgenstern utility function; given this, if reasonable people find themselves in PD situations — that is, if their utility functions ordered the outcomes in a way that defines the PD — they too would follow the dominant "defect" strategy. The difference between simply rational agents and those who are also reasonable is not that they would behave differently in Prisoner's Dilemmas, but that reasonable people are more successful at avoiding the Prisoner's Dilemma and tend to play more cooperative games.

2. The Prisoner's Dilemma and Utility

2.1 The all-too-familiar story behind the prisoner's dilemma goes like this. Two suspects, Alf and Betty, have been arrested by the police. The police have enough evidence to convict both on a relatively minor charge. If convicted of this charge — and the police can obtain a conviction — each will get two years in prison. The police, however, suspect that Alf and Betty acted together to pull off a bigger crime but the police have inadequate evidence to convict them of that crime. They make the following offer to Alf (the same offer is made to Betty). "Alf, turn state's evidence against Betty, and we'll let you go free; we'll demand the maximum penalty for Betty, so she will get 12 years. Of course if Betty confesses too, we're not going to let you both go free: you'll each get 10 years. However, if you keep quiet and she confesses, we'll let her go free, and you will be the one to get 12 years. But if neither of you confess to the bank job, we won't have enough evidence to prosecute. We will then proceed with the lesser charge, and you'll get two years each." Figure 1 displays their problem in terms of years in jail; Alf's "payoffs" (time in jail) are depicted in the lower left of each cell, Betty's in the upper right.

		Betty	
		Keep Quiet	*Confess*
Alf	*Keep Quiet*	**2** **2**	**0** **12**
	Confess	**12** **0**	**10** **10**

Figure 1: PD in Terms of Years in Jail

Alf reasons: "If Betty confesses, and I keep quiet, I'll get 12 years; if Betty confesses and I confess too, I'll get 10 years; so I know one thing: if Betty confesses, I better confess too." What if Betty keeps quiet? Alf reasons: "If Betty keeps quiet and I keep quiet too, I get 2 years; if Betty keeps quiet and I confess, I go free. So if Betty keeps quiet, I do best by confessing." But now Alf has shown that confessing is a *dominant strategy*: no matter what Betty does, he does best if he confesses. And Betty will reason in a parallel way; she will conclude that no matter what Alf does, she does best by confessing. So they will both confess, and get 10 years. Hence the (sole) equilibrium outcome is strongly Pareto-inferior to the non-equilibrium outcome {keep quiet/keep quiet}.

2.2 This, however, is simply a story in terms of jail time. We have simply assumed that the players want to stay out of jail, and that their utility functions are monotonic with minimizing jail time. In order to really get the result that the rational thing for them to do is to confess, we need to say something about their *preferences* over outcomes. We can generate an ordinal utility function for any person in terms of his preference rankings for the different outcomes if his rankings satisfy the standard conditions of completeness, asymmetry of strict preference, symmetry of indifference, reflexivity and transitivity.[1] Ordinal utility functions map rankings of outcomes on to numbers. Let us assume that most preferred outcome is mapped on to the highest number, the next preferred to a smaller number, the next to a yet smaller number and so on. The sizes of the differences, or ratios between the numbers, provide no additional information.

Assuming that in both of their preference orderings less years in jail are preferred to more years (and, remember, our ordinal scale is one in which *larger* numbers designate *more* preferred outcomes), we get Figure 2.

		Betty	
		Keep Quiet	*Confess*
Alf	*Keep Quiet*	**3** **3**	**4** **1**
	Confess	**1** **4**	**2** **2**

Figure 2: General PD Form in Ordinal Utility

Figure 2 is the general ordinal form of the prisoner's dilemma. Each ends up with his/her third ranked outcome (utility 2), yet {keep quiet/keep quiet} would give each his/her second choice (utility 3). Thus even though there is a strongly Pareto-superior outcome (i.e., one that is preferred by each) they cannot achieve it. Ordinal utility only allows us to distinguish more and less preferred outcomes; rather than {4, 3, 2, 1} we could have used {1000, 999, 4, 1}, which would give precisely the same information. If we wish to (roughly speaking now) get some idea of the relative preference distances between the outcomes (again, roughly how much more one thing is preferred to another),[2] we then can generate cardinal utilities, using some version (there are several) of the standard von Neumann-Morgenstern axioms. On one accessible view, four further axioms are required.[3] The key to this approach is to assume certain preferences over lotteries (risky outcomes), and then confront agents with lotteries involving their ordinal preferences. Their ordinal preferences *over the lotteries* allow us to infer a cardinal scale (or, rather a set of such scales, since the results are unique only up to linear transformations). This is an incredibly powerful idea, for it generates a cardinal utility measure from series of ordinal preferences. We can define a cardinal utility PD as in Figure 3:

		Betty	
		Keep Quiet	*Confess*
Alf	*Keep Quiet*	x x	1 0
	Confess	0 1	y y

Where $1>x>y>0$
Figure 3: The General Cardinal Form of a PD

Figure 4 gives one example of how such cardinal utilities might come out.

		Betty	
		Keep Quiet	*Confess*
Alf	*Keep Quiet*	**.85** **.85**	**1** **0**
	Confess	**0** **1**	**.1** **.1**

Figure 4: A PD in Cardinal Utility

So Alf and Betty reason themselves into an outcome that, on each of their cardinal scales, each ranks as giving him/her .1 out of 1, whereas the {keep quiet/keep quiet} outcome would give each .85 out of 1.

3. Can the Utility of Being a Reasonable Person Be Included in the Game?

3.1 The idea, then, is that purely rational, means-end oriented, agents will sometimes find themselves in PDs. If, though, they reasoned in another, more cooperative, way they could avoid the Pareto-inferior outcome. Recall that Rawls describes a reasonable person as one who has a "desire to engage in fair cooperation as such." Thus we might say that while purely rational people will only be cooperative when doing so is the path to largest payoff, rational and reasonable players will gain intrinsic utility from taking the cooperative moves (they have a desire to be cooperative "as such"). Of course, Rawls also adds that a reasonable person is one who is concerned with conditions of *fair* cooperation. So we might say that a reasonable person intrinsically values taking the fair cooperative move. This means that she will choose the cooperative move when others do so as well, since it is not fair to demand that anyone be an unconditional cooperator; proposing that others cooperate on those terms — "You cooperate no matter what I do"— is not reasonable. Assume, then, that the players are reasonable insofar as they have a preference to be conditional cooperators: people who cooperate when others do. Apart from the payoffs to which such a cooperative stance might lead, each has a preference to be a conditional cooperator rather than a person who seeks to gain by unilateral defection. So, someone might be tempted to say, rational and reasonable people might cooperate in a PD.

3.2 Now the obvious response by a traditional game theorist is to insist that all the utility that is relevant to the game must be built into the game. Suppose, then, that each player values being seen as a cooperative person, but not as a sucker (that is, each puts intrinsic value on cooperating when the other cooperates, but not on cooperating when the other takes advantage of one). Adding .2 extra units of utility for performing the cooperative act, we get the game in Figure 5:

		Betty	
		Keep Quiet	*Confess*
Alf	*Keep Quiet*	**.85 (+.2)** **.85 (+.2)**	**1** **0**
	Confess	**0** **.1**	**.1** **.1**

Figure 5: The Transformation of a PD into an Assurance Game by Adding the Utility of Cooperative Action

In this game {keep quiet/keep quiet} is in equilibrium (if one player keeps quiet, the other cannot improve his/her total utility of 1.05 by confessing). Unfortunately, like so many attempts to "solve" the Prisoner's Dilemma, we have done so by converting it into another game, in this case the "Assurance Game": the utility of the outcomes no longer conforms to Figure 3, the general form of the Prisoner's Dilemma. In the Assurance Game there are two equilibria: {keep quiet/keep quiet} and {confess/confess}.

Important decision theorists, however, insist that this is not the proper analysis. It looks now as if in Figure 5 we stipulate that *cooperating is the path to the most utility*, but this seems to miss the idea that in some sense our reasonable and rational cooperators could gain by defecting. Thus Amartya Sen argues that those in a Prisoner's Dilemma who constrain their pursuit of the best payoffs act "as if" they are in an assurance game, but they are not apparently really playing one (Sen, 2004: 218). Robert Nozick agrees. He insists that some utility cannot be integrated (as I have done in Figure 5) into the payoffs in the game. Sometimes, Nozick argues, an act's utility "is not determined solely by that act. The act's meaning can depend upon what other acts are available with what payoffs and what acts are also available to the other party or parties" (1993: 55). Thus an act's utility "may depend on the whole decision or game matrix. It is not appropriately represented by some addition or subtraction from utilities of consequences *within* the matrix" (1993: 55, emphasis in original). So the idea seems to be that a certain utility may depend not just on the value of a consequentially resulting state of affairs, but on the entire game, including what other options are available to both players. Nozick insists that this cannot be captured within, as we might say, any single cell but depends on the relation between the cells (the "whole game matrix").

3.3 Sen, then, thinks that reasonable cooperators are playing "as if" they were in assurance game; Nozick believes that we need to distinguish payoffs which are simply the *results* of an action from payoffs that depend on players having confronted certain options in the course of the game. Is there any way to analyze the game in Figure 5 that makes sense of these intuitions? Nozick focuses on the "matrix" — the strategic representation of the game. However, as soon as we become concerned about the information available at different points in a game (which player had what options), the strategic form is inappropriate, and we should consider the extensive form of the game. Figure 6 provides the extensive form of our game in Figure 5 — "Transformation of the PD into an Assurance Game by Adding the Utility of Cooperative Action Game" (hereafter the "Reasonable Cooperators Game").

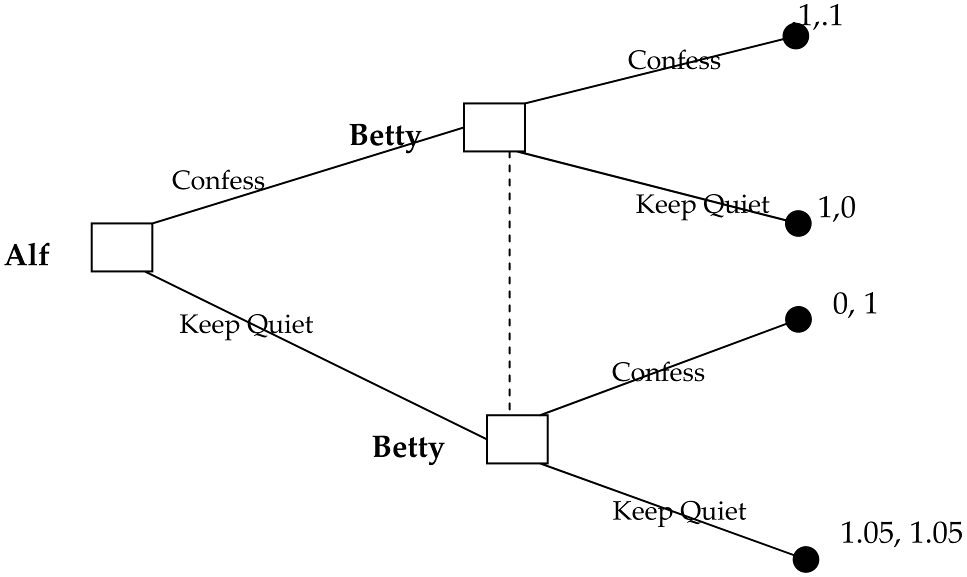

Figure 6: The Reasonable Cooperators Game in Extensive Form

Squares indicate decision nodes, filled dots are terminal nodes that indicate the end of the game, or the payoffs of the game (the utilities are given in cardinal numbers, first Alf's, then Betty's, at each terminal node). The advantage of the extensive form is that at each node we can identify the information sets available to the players. We can specify that in games of "perfect recall" information sets include knowledge of the prior moves of both oneself and the other player made at each node. The extensive form builds into games the order of the moves; in Figure 6 Alf makes the first move. However, in prisoner's dilemma-like games, the moves are simultaneous. This feature is accommodated in Figure 6 by the dotted line connecting Betty's decision nodes; she must make her first move without knowing which node she is at (her information set at this node is thus not a singleton, as she does not know which of the two nodes she occupies). Consequently, the same game could be displayed with Betty having the first move, and Alf making the second move with his information set incomplete in a similar way. Now when we think outside the box (strategic form) of the game in this way, we see how the utilities at the terminal nodes can be affected by information about what nodes the players have passed through. Alf's utility of 1.05 (the same holds for Betty) is produced (partly) by his knowledge that at choice points (nodes) where he might have ratted on Betty and she might have ratted on him, they *both* chose not to, and instead took a more cooperative path.

Is it legitimate to interpret a game in this way — where the utility of the terminal nodes is dependent on the players' knowledge of what decisions have been made at earlier nodes, and what this tells them about each other? Consider the game of chicken in Figure 7:

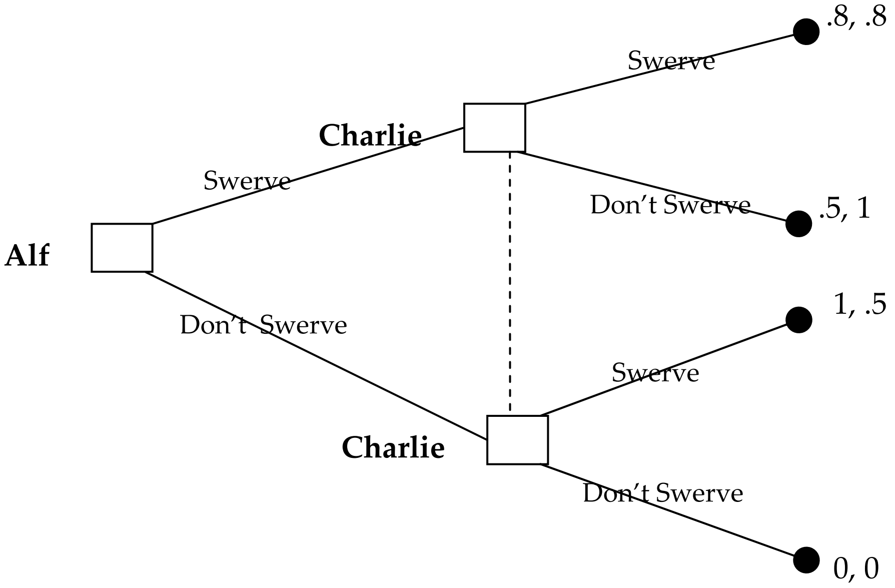

Figure 7: Chicken

Figure 7 is a familiar textbook game. Its standard name comes from the teenage game in the 1950's, in which two teenage boys (or, as Bertrand Russell put it, "youthful degenerates") drove toward each other with the pedal to the metal, and the first one who swerved was "chicken." So the winner gets 1 out of 1 if he keeps driving straight and the other swerves (say the swerver gets .5 out of 1); if both swerve their reputations take a bit of a hit (say their payoff is .8 out of 1). If, on the other hand, neither swerve they both take a much bigger hit and crash (0 out of 1).

What is seldom appreciated in the textbook rendition is how much of the intuitive description depends on the knowledge by each player at the end of the game what turns at each node he has or has not taken and what turns at each node the other has or has not taken. Like the Reasonable Cooperators Game, the intuitive account of the game is crucially about what sort of person one is and this is determined by the choices one made at each node. So far from this utility — derived from knowledge from each other's choice set — being "outside the game," it is most of the game (along with the disutility of surviving or being killed). To better see this, suppose that it was discovered that one of the cars was made in Sweden and had a safety auto-swerve device such that, when another car was approaching, at a distance of 30 feet the car automatically turned away. Although in *some sense* (see §4.3) the consequences would be "the same" as if one player chickened (we get to a "swerve, didn't swerve" terminal node), the payoffs would change, since the swerving was not the result of the other player making a chicken choice at one of the nodes.

Often all games with the payoffs ordered as in Figure 7 are considered the game of Chicken. I believe this is wrong. Consider a different game of "Chicken" drawn from Dennis Mueller (2003: 16). Suppose Alf's goat wanders into Betty's garden and eats her veggies while Betty's dog wanders on to Alf's property, scaring his goat so that it does not give milk. A fence would be a public good between the two of them. Assume that each would benefit by unilaterally building the fence (each would be better off building the fence alone than not having one) but, of course, each would prefer that the other build the fence. So each has the following ordering: (1) the other builds, (2) we split the cost (3) I build; (4) neither builds. We get the following game in ordinal utility (4=best).

Betty

		Builds	*Doesn't Build*
Alf	*Builds*	3 3	4 2
	Doesn't Build	2 4	1 1

Figure 8: "Chicken" in Providing a Public Good

This game has the same ordinal strategic representation as the game of Chicken of Figure 7, but it is crucially different. Here the payoffs in no way depend on having traveled through certain choice nodes: Alf's and Betty's payoffs are determined exclusively by the resulting state of the world (whether a fence is built or not, and who pays) and they get no payoff at all from knowing that the other party "chickened" out.

3.4 Figures 7 and 8 represent different games *even though the orderings of the payoffs are identical*: the decision trees for the two games have different properties. To see the importance of this, compare the following two choice situations confronting Alf:

(1) *Betty's Reward*: Betty says to Alf "I don't want you to go the football game this afternoon or go drinking with your friends tonight. Forgo both and you will get a kiss from me this evening."

(2) *Alf's Trek to Betty*: Alf wants to see Betty this evening and get a kiss (if he turns up at her door, he *will* get a kiss), but on the way will confront choices between seeing a football game or continuing on, and then he will confront the choice between going into the bar or continuing on to see Betty. If he goes to the football game he will be too late to drink beer or see Betty; if he goes drinking he will also be too late to see Betty.

Figures 9 gives Alf's decision tree for Betty's Reward and Figure 10 gives it for Alf's Trek.

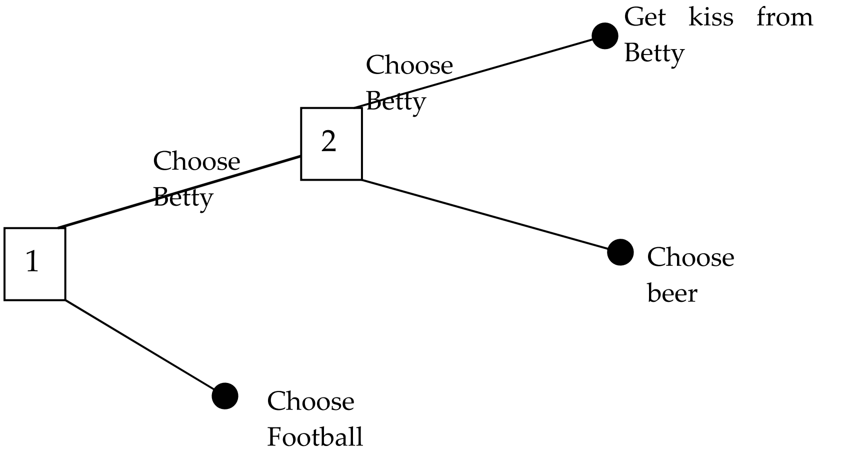

Figure 9: Alf's Decision Tree for Betty's Reward

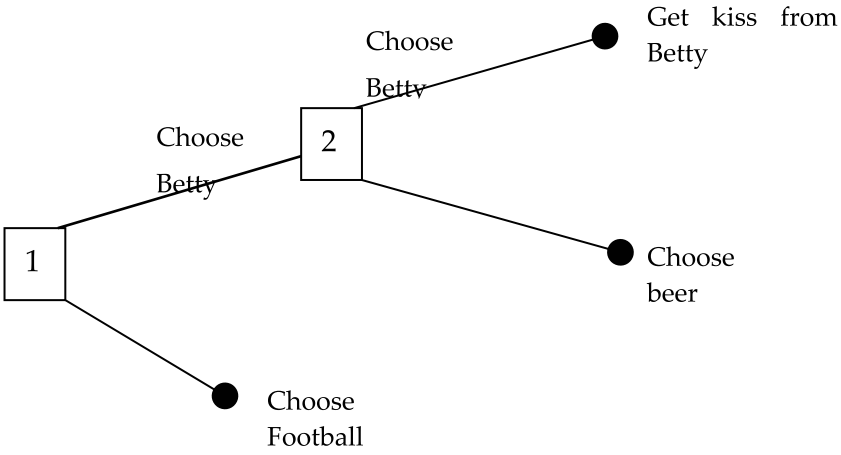

Figure 10: Alf's Decision Tree for Alf's Trek

These trees look identical; however, they differ in a crucial respect. The tree in Figure 10 is *separable*, in the sense that if we truncate the tree, starting at node 2 rather than 1, this separated part of the tree is the same as it was when it was part of the larger tree (McClennen, 1990: 120ff). In Figure 10 Alf has the exact same choice open to him at node 2, whether we begin the tree at node 1 or node 2: Betty's kiss or beer. But not so in Figure 9: the payoffs there depend on passing through both nodes, so it makes no sense to truncate the tree.[4] Alf cannot start at node 2. Clearly in *some* decision trees the payoffs are necessarily conditional on confronting a series of choices and so, in such cases, a separable game cannot be started at a node that does not include one of the choices. Trees are not always simply "access routes to prospects" (cf. McClennen, 1990: 120). Because in a genuine game of "Chicken" as well as the "Reasonable Cooperators Game," the payoffs depend on proceeding through certain nodes, the players' decision trees will have similar difficulties with separability, since the structure of the decision tree is part of the payoffs (Hammond, 1988: 26).

3.5 Nozick and Sen are (of course) right. Just because two games have the same payoffs — the games are the same in their strategic form — they may nevertheless be different games in their extensive form. The decision trees for the player's may have different properties even though they lead to the exact same utilities. But this by no means justifies the idea that somehow the utilities gained by the players from their knowledge about what moves have been made cannot be integrated into the payoffs of the terminal nodes (or cannot be included "in the matrix"). All the utilities at stake in a game are part of its payoffs. The difference between a standard Assurance Game and the Reasonable Cooperators Game is not that there are some extra payoffs lurking somewhere outside the matrix in the latter. The difference is not in the payoffs at all, but in the characteristics of the game, which make the payoffs depend on passing through certain nodes. Nevertheless, reasonable agents thus described find themselves in strategic interactions that are versions of the Assurance Game; and so understanding such games looks crucial to grasping how reasonable and rational others will interact (Skyrms, 2004). It is not discovering how to "cooperate in PDs," but finding the cooperative equilibrium in these types of Assurance Games, that explains the emergence of cooperation of rational and reasonable individuals.

4. Self-Sacrifice Arguments Against Integrating All Utility into the Payoffs

4.1 My claim, then, is that in a game everything of normative relevance for choice — "even the structure of the decision tree itself"— is part of the consequence domain (Hammond, 1988: 26). The utility at the terminal nodes sums up all the normatively relevant considerations. In some ways, Sen agrees that the moral person is a maximizer, but in other ways, Sen argues, she isn't. After all, she derives utility from taking a *less* attractive option. Thus, says Sen:

> A person's preferences over *comprehensive* outcomes (including the choice process) have to be distinguished from the conditional preferences over *culmination* outcomes *given* the act of choice. The responsibility associated with choice can sway our ranking of our narrowly-defined outcomes (such as commodity vectors), and choice functions and preference relations may be parametrically influenced by specific features of the *act* of choice (including the *identity* of the chooser, the *menu* over which the choice is made, and the relation of the particular *act* to behavioral social norms that constrain particular actions) (2002c: 159).

Sen distinguishes the "comprehensive" outcome (which can include the utility of the choice process) from the distinct state of affairs that is produced by a choice, the "cumulative" outcome. Insofar as part of the outcome derives from what it shows about one or the options confronting one, this is part of the comprehensive, but not the cumulative, outcome.

Sen has in mind cases in which the utility of the states of affairs depends on the fact that one passed up what looked to be a more attractive option.

> You arrive at a garden party, and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you to that chair. However, if the matter is left to your own choice, you may refuse to risk it. You select a "less preferred" chair. Are you still a maximizer? Quite possibly you are, since your preference ranking for choice behavior may well be defined over "comprehensive outcomes," including choice processes (in particular, who does the choosing) as well as outcomes at culmination (the distribution of chairs).
>
> To take another example, you may prefer mangoes to apples, but refuse to pick the last mango from the fruit basket, and yet be very pleased if someone else were to "force" that last mango on you (2002c: 161, footnote omitted).

Now, on the face of it, this sort of chooser seems to act irrationally. Suppose one is confronted with the option {mango, apple}; given one's preference not to take the last mango, one will choose an apple. But now suppose that one is confronted with the set {mango, mango, apple}. Now one will pick a mango. But our last mango refuser will violate what many take to be basic axioms of consistent rational choice — the contraction and weak expansion properties. According the *contraction* property, if *x* is chosen from the entire set *S*, it must be chosen from all subsets of *S* in which *x* is included. Our polite mango refuser violates this by selecting a mango from the set {mango, mango, apple} but an apple from the subset {mango, apple} (Anand, 1993: 56-58). Our chooser will also violate the *weak expansion* principle: if an option is chosen from each of two subsets, it must still be chosen when the sets are combined.[5] Suppose our person is confronted with two sets {apple, apple, mango} and {apple, mango}. Because she will not take the last mango, she will chose {apple} from the first set and {apple} from the second. But if we combine the two sets to get {apple, apple, apple, mango, mango} she will choose a mango, thus violating the weak expansion principle.

Suppose, as I think is clearly the case, that having such preferences is rational, and so we want to allow for them in an account of consistent choice; it may look as if we must follow Sen in developing new axioms of rational choice, distinguishing choices from menu-independent sets (where the contraction and weak expansion principles may hold without modification) from axioms of choice involving options, like the choice of our mangoes, that are chooser or menu-dependent.[6] Thus when Sen argues that conditional cooperators act "as if" they are in an assurance game, the idea is that the best modeling of their utility function is that they maximize their goals subject to a self-imposed restriction to a certain menu. That is, rather than (as I have argued) building into the cooperative person's utility function their preference for cooperative acts, Sen argues that we can better capture their deliberations as constraining their maximization behavior to a certain subset of the options.[7] So on Sen's view we model the person as first restricting her action options by identifying a "permissible" subset of her options "reflecting *self-imposed* constraints, and then seeking the maximal elements" within that remaining set (2002c: 189ff).

4.2 This proposal fits well with some understandings of the reasonable, *viz*., in which deontological principles function as side constraints. However, the idea of a self-imposed menu constraint does not obviously capture deontological *requirements*, which do not function primarily as constraints on maximization. Principles that require positive action are not easily interpreted as menu constraints. If the principle requires performance of *x* out of the set {*x*,*y*,*z*}, it doesn't look as if the principle is one of constrained maximization: it dictates a choice. To be sure, it might be said that in this case the principle constrains the option set to one item and then that is maximized, but it is not clear what it means to appeal to maximization to determine what option to select from the restricted menu when the option to be acted upon has already been selected. Perhaps we could still make sense of this. After all, there is typically more than one way to fulfill a principle, and perhaps we can maximize when selecting from the various ways. Yet this too seems normatively charged; some theories may well instruct us not to appeal to our own goals when deciding among alternative instantiations of the principle, but instruct us to consult the spirit or *telos* of the principle.

4.3 I am not convinced that we should accept Sen's complication of decision theory to model this type of "sacrificing" choice. The polite last-mango refuser only violates the principles of consistent choice (contraction and weak expansion) if each choice is viewed as ranging over enjoyable food items. If Betty has stable preferences, and is simply picking the more tasty fruit, and if Betty chooses a mango when presented with the choice between a mango, an apple and another apple, it is perplexing indeed if she then chooses an apple when confronted with the choice between a mango and an apple. It looks quite irrelevant that the first time her set included an extra apple. But, of course, the problem arises just because the relevant description changes: at one point Betty is choosing simply on the grounds of "which fruit would I like the best?" and at the other time the relevant description is "Should I choose the one I like the best or be polite, knowing that Alf loves mangoes?" If Betty has reasons according to which, in cases like this, being polite is more important than an enjoyable fruit fest, then she is simply acting on her total set of preferences and there is no inconsistency.

This raises the difficult issue of "framing" and whether Betty's choices violate what Kenneth Arrow calls "extensionality":

> The cognitive psychologists refer to the "framing" of questions, the effect of the way they are formulated on the answers. A fundamental element of rationality, so elementary that we hardly notice it, is, in logicians' language, its *extensionality*. The chosen element depends on the opportunity set from which the choice is to be made, independently of how it is described (1982: 6).

Now if, when the *same option* is described in different ways a person's utility changes, extensionality (or invariance) is violated (Tversky and Kahneman, 2000: 211.) Is Betty just framing the same choice differently? That is, can we say that she *really* has a choice between eating a mango and an apple, but she responds to different descriptions and so changes her preferences? This gets us into complex issues in the philosophy of social science, regarding the intentionality of actions. What I think is clear, though, is that there is no such thing as a set of brute action options that is independent of the descriptions (intentional states) of the choosers. Are Betty's true options: a mango or an apple to eat, a soft object or a hard one, a dull-surfaced object or a shiny-surfaced one, the superior piece of fruit to throw at a disliked political speaker, the superior fruit to put on the teacher's desk, or between being rude or being polite? One of the hopes of revealed preference theory, with its behavioral underpinnings, was that we could describe an unambiguous "choice behavior" that had no reference to the chooser's intentional states, and so her descriptions of what she is up to. But this behaviorist project failed: action is inherently intentional. So "framing" cannot simply be understood in terms of different descriptions of the "same" option, for what is the "same" option depends on the relevant description. Sen, I think, agrees: framing explains inconsistent choices, but as he sees it, Betty's fruit choices do not really seem inconsistent. (2002c: 168n). A full account of framing, and its relation to a plausible version of Arrow's condition of "extensionality," must involve a notion of *irrelevant* differences in description or a criterion of choice inconsistency.[8]

5. Decision Theory: *Pro Tanto* or All-things-Considered Considerations?

5.1 So how do we model in decision theory and game theory a person's choosing as in some sense "the best" an action that is nevertheless a self-sacrificing choice, so she (in a sense) loses utility? My suggestion is that we do not explicitly do so. Of course our intuitive description of a game can include these: we can say that in the Reasonable Cooperators Game each person has adopted a moral principle not to cheat if the other cooperates, and this can be seen as a sort of sacrifice, but it is not modeled in the game itself.

To better see the view I am espousing, contrast three conceptions of decision theory. The first I have mentioned and will put aside. Decision theory was, as I said, originally presented as a theory of consistent *choice behavior*, where it was hoped that this might entirely avoid relying on mental states. I have indicated why this aspiration was misconceived.

5.2 This leads to the second conception. Decision theory is crucially concerned with how a person's preferences over states of affairs translate into her preferences over action-options. We need to suppose, most basically, that a person can rank the possible relevant states of the world in terms of her normative criteria, whatever they are. Let us call the *consequence domain* the ordering of possible outcomes: the ordering sums up everything relevant in the person's set of normative criteria to ranking the states of the world that might obtain. Now suppose a chooser confronts a set of *action options*; she will rank the action-option highest that is associated with the highest ranked outcome in the consequence domain. This is what we mean by saying that she has more reason to choose that act: given her entire set of normative criteria, doing that act is preferred to all the other alternatives. Thus, her preferences over outcomes (the consequence domain) determine her preferences over action-options. We can think, then, in terms of mapping the ordering of outcomes on to the action-options set, producing an ordering of action-options (Morrow, 1994: 17).

The power of decision theory is that modest principles of consistency and transitivity of preference allow us to construct a mathematical representation of a person who consistently acts on her best reasons — i.e., chooses higher- over lower-ranked options and has a complete ordering of outcomes; for cardinal representations additional and somewhat more contentious principles are required, but they too are pretty intuitive. This mathematical representation allows us to depict consistent choices for higher- over lower-ranked options as maximizing a utility function. Decision theory then formalizes a person's *all-things-considered considerations* in favor of action options based on her ranking of outcomes. It is crucial to stress that decision theory simply does not maintain that anyone *seeks* to maximize utility — that idea is a remnant of utility *qua* hedonism. Acting in a way that maximizes utility models choices that are consistent with one's ordering of action options based on one's ordering of outcomes; maximization of utility is not itself a goal.

5.3 It is absolutely fundamental to realize that there is no reason whatsoever to suppose that a person's background set of evaluative criteria must produce an ordering of outcomes that ranks states of affairs simply in regard to how well a person's *goals* or *welfare* are achieved (Cf. Morrow, 1994: 17). Although decision theory distinguishes acts from outcomes (or consequences), and holds that the ranking of acts is determined by the ranking of outcomes, we should not confuse this sort of decision-theoretic consequentialism with the moral theory of consequentialism or the theory of instrumental action (Anand, 1993: 84n). Among an agent's background evaluative criteria may be to conform his actions to the moral principle to "tell the truth when under oath." Suppose, given one's evaluative criteria, one ranks at the top the outcome "I tell the truth under oath at the trial today." Given this, the action of telling the truth under oath has "high utility" — that is, the action one has most reason to perform. S.I. Benn has shown that deontological requirements can be modeled in this way (1988: ch.3).

It is a mistake, albeit a common one, to see decision theory as a theory of instrumental action.[9] Decision theory allows us to model choice based on one's notion of the overall ordering of outcomes based on one's evaluative criteria or, we might say, one's best reasons — whatever they are. To be sure, if one also claims that all reasons are reasons to achieve one's goals, *then* decision theory does indeed model instrumental reasoning, but only because one's practical reasons have been limited to goal-seeking ones. If one's practical reasons include being a fair cooperator and these reasons lead to ranking outcomes in ways the meet the basic utility axioms, then a person acting on her best reasons can be modeled as maximizing a mathematical cardinal function. Gary E. Bolton (1991) has done this, building into players' utility function (along with the goal of getting money) a concern for fairness *to themselves* (i.e., that the player is himself treated fairly); moreover, Bolton provides experimental evidence that this model predicts choices in bargaining games.

5.4 Sen dissents from this way of modeling the action: he advises us to distinguish actions that follow from "adhering to a deontological principle" from those that are "actually 'preferred'" (20002c: 191). The idea is that an obligation that requires one to act in a way that sets back one's goals or welfare (perhaps my best friend will be convicted if I tell the truth under oath) is not an action I "prefer" to perform. Here Sen is pushing the idea of "preference" closer to its ordinary meaning of "liking," where one can rationally do what one does not prefer ("I had reason to do it, but I sure did not prefer it.") (Benn and Mortimore, 1976: 160-161). R. Duncan Luce and

Howard Raiffa hint at a similar interpretation of preferences when they refer to them as “tastes”: if preferences are tastes, then it is surely wrong to describe a Kantian as one who has a “taste” for justice.[10] Given this, we can see that Sen, Nozick and those who resist integrating the reasonable into normal utility functions seek a third conception of decision theory: one that does not simply model the relation of *all-things-considered* orderings of outcomes to choice and action, but endeavors to model our *pro tanto* reasons and how we structure them to arrive at all-things-considered rankings and choice. Thus, as we have seen, Sen models deontic constraints differently from goal maximization. Sen (and Nozick) seek a decision theory that *models rational deliberation and its relation to choice and action*. If we have different types of reasons, then the decision theoretic model should distinguish these to stay truer to the phenomenon of rational deliberation, choice and action.

There is nothing erroneous about transforming decision theory from an account of all-things-considered rational choice to model *pro tanto* considerations as they enter into all-things-considered choice. In some ways microeconomics does this: it models preferences over consumption of goods (where the preferences are subject to further conditions, such as decreasing rates of marginal substitution) subject to budget constraints. And we have long been familiar with metapreference analysis, which supposes that a rational agent has first-level preferences and also preferences about these preferences (and so on up levels). Sen, consistent with his general approach, calls such “metarankings…an analytically tractable concept” that has been “practically important” (2002a: 12).

Nevertheless, there are reasons to think this development of decision theory into a theory of choice based on structured *pro tanto* reasons, while in many ways interesting, is ill-advised. Not only does decision theory become increasingly complex, but more importantly, it becomes tied to different accounts of types of reasons. Sen, we have seen, develops a way of modeling deontic reasons as side constraints, but it is not clear that he adequately models deontic requirements. Nozick argues that the game theory cannot integrate the utility associated with "symbolic" cooperative reasons, but there is good reason to doubt whether they are reasons at all (Nozick, 1993: 54-55; Pincione and Tesón, 2001; Gaus, 2002). Just what reasons we have, and how they are to be distinguished, is philosophically highly contentious. Now we might develop models for each of these: a virtue decision theory, purely instrumentalist decision theory, a side-constraint decision theory, a deontic requirement decision theory, and egoistic decision theory, and so on. But as we do so, decision theory loses its appeal as an ecumenical theory that can relate a person's (overall) rankings of outcomes to choosing actions, and so understanding how people with differently ordered outcomes may rationally interact. Furthermore, if we follow Nozick's (and Sen's) lead, and see games such as the Prisoner's Dilemma as only about how the players' goal-related reasons would instruct them to act, the games become difficult to interpret, and, crucially, *the overall rational course of action turns on reasons not identified in the game's payoffs*. Thus we have "solutions" such as Nozick's to the Prisoner's Dilemma which really turn on the claim that the game is under-described since the reasons that tilt the balance to cooperation (that lead the players to order {keep quiet/keep quiet} above {I confess, the other keeps quiet}) are not "in the matrix." Essentially, games are described in terms of partial utility — *pro tanto* reasons — and the claim is that, *if* these are one's only reasons, then we will behave as the game predicts.

7. Conclusion: Morality and Utility Theory

I do not want to claim that every normative criterion can be accommodated by cardinal utility theory without complications.[11] Cardinal utility theory is full of complications. Take the straightforward problem of the deontologist who places absolute weight on adhering to a moral principle. So *x*, the world in which he abides by the principle, is best. Suppose that the next best outcomes (worlds) are *y* and *z*. One of the axioms of cardinal decision theory, continuity, says that there must be a lottery *L* in which he is indifferent between *y* and a lottery that gives him *p* probability of *x* and 1-*p* of *z*.[12] But our absolutist prefers *x* for any probability over zero. As Luce and Raiffa (1957: 27) acknowledged, some choices may not be continuous. To use their example: even if we all agree that $1≻1¢≻death, not too many people are indifferent between 1¢ and a lottery with chance *p* of $1 and a 1-*p* chance of death. It might be thought that the whole idea of a lottery over prizes makes no sense for a deontic theorist who always has total control over his action, and so his "prizes." The absolutist is acting under certainty, not risk, so the lottery axioms are inappropriate.

But note that these problems concern cardinal utility measures. An absolutist still can have a complete, reflexive and transitive preference ordering (at least, as long as he has only one absolute). As Rawls notes, a strict lexicographic preference-ordering prevents formulating a cardinal utility function (1996: 332n). And if we see deontologists (qua deontologists) as always acting under certainty (which I think is erroneous), then indeed we will not employ expected utility accounts, which model choices under risk. The important point, though, *is that these sorts of worries cannot show that decision theory is about instrumental reasoning (or is consequentialist in any interesting sense)*: they are objections to the lottery axioms and the development of cardinal utility. I have given games in both ordinal and cardinal utility: the difference between them is the amount of information conveyed by the utility numbers about the relation between the ranked outcomes. The difference is not that cardinal utility commits us to instrumentalism but ordinal utility does not. If so, these problems with modeling some sorts of deontic choices may be barriers to developing a cardinal utility scale modeling such choices, but this by no means shows that deontic choices cannot be modeled in decision theory because it is "consequentialist."

However, placing infinite weight on a moral consideration is an extreme position indeed. A person who chooses on the basis of pluralistic reasons to act, acknowledging both goal-oriented, means-end reasoning, and moral reasons that place intrinsic value on doing certain sorts of acts, and who never gives infinite weight to any reason, can be modeled according to a cardinal utility function (Benn and Mortimore, 1976: 185-186). Although those who insist that the moral or the reasonable cannot be integrated into utility theory are apt to see themselves as followers of Rawls, this was not, in fact, his view. "From a purely formal point of view, there is nothing to prevent an agent who is a pluralistic intuitionist from having a utility function" (1996: 332n).

Philosophy Department
University of Arizona

Notes

Earlier versions of this paper were presented to the 2005 Fagothey Philosophy Conference, Santa Clara University and the Universidad Torcuato Di Tella, Buenos Aires. My thanks to all participants for their comments, questions, and objections.

[1] C*ompleteness:* For every option (x,y) it must be the case that either x is preferred to y, y is preferred to x, or y and x are indifferent. Let us use "$x \succ y$" for "x is preferred to y"; "$\mathrm{x} \sim y$" for "x is indifferent to y" and "$x \succsim y$" for "x is either preferred to y or x is indifferent to y." So for all (x,y): $x \succsim y \vee y \succsim \mathrm{x}$; *Asymmetry*: not $(x \succ y$ & $y \succ \mathrm{x})$; indifference is *symmetric*: if $x \sim y$ then $y \sim x$; *Reflexivity:* $x \succsim x$; *Transitivivity*: $x \succsim y$ & $y \succsim z \rightarrow x \succsim z$.

[2] If we wish to be extremely careful, we will restrict ourselves to saying that all these "von Neumann-Morgenstern" utililities tell us are a person's preferences between lotteries or gambles, and so what he will do in situations that involve risk — where the agent does not know for certain what outcome-consequences are associated with his action-options, but can assign a specific probability p that a certain action option α will produce a certain consequence. See Morrow, 1994: 34. See further §§5.2, 7.

[3] *Continuity*: For all (x,y,z) where $x \succsim y$ & $y \succsim z$ there must exist some probability p such one is indifferent between y and a lottery L that gives one p chance of x and $1\text{-}p$ chance of z; *Better prizes:* if (i) we are confronted with lotteries L_1 over (w,x) and L_2 over (y,z); (ii) L_1 and L_2 have the same probability of prizes; (iii) the lotteries each have an equal prize in one position — (w~y) or (x~z); (iv) they have unequal prizes in the other position; then

(v) if L_1 is the lottery with the better prize, then $L_1 \succ L_2$; if neither lottery has a better prize, then $L_1 \sim L_2$; *Better chances*: if (i) confronted with a choice between L_1 and L_2, and they have the same prizes; (ii) if L_1 has a better chance of x (recall that the lotteries are between x and z, where $x \succeq y \succeq z$), then $L_1 \succ L_2$; *Reduction of compound lotteries*: If the prize of a lottery is another lottery this can always be reduced to a simple lottery between prizes. I follow Dreier, 2004. But see also Hampton, 1998: ch. 7; Luce and Raffia, 1957: 23-31.

[4] As Peter Hammond (1988: n. 4) notes, "a decision tree can hardly include, as a partial consequence, regret at missing an opportunity to have consequence *y*, unless there was an opportunity in the past to have had *y*." Hammond argues, however, that his continuity principle over choices may still apply. In any event, note that denying separability does lead one at any node to choose what, from that node onwards in the tree, would be a suboptimal outcome. Thus the core of modular rationality is retained.

[5] Weak expansion is crucial to the idea of path-independent choice. I call it the "weak expansion" principle as it less demanding than Sen's β+ property. See Riker, 1988: 132ff; Mueller, 2003: 152-153; Craven, 1992: 63ff.

[6] Sen's argument is complex. He argues for a notion of maximization that is distinct from optimization, which itself has to drop consistency conditions. I cannot go in to these matters here. See Sen, 2002c: 184n.

[7] For a similar approach see McMahon, 2001. I examine McMahon's proposal in Gaus, 2003.

[8] Arrow himself refers (1982: 7) to people being moved by "irrelevant" events. On justifying distinctions between preferences, see Broome, 1991.

[9] For an extremely insightful if contentious analysis, see Hampton, 1988: ch. 7. David Gauthier (1986: ch. 2), makes the error of conceiving of decision theory as instrumental. Morrow presents a typical though erroneous interpretation: "Put simply, rational behavior means choosing the best means to gain a predetermined set of ends" (1994: 17).

[10] They did acknowledge (1957: 21) that this is a very rough interpretation. Cf. Hampton 1988: 239-240n. So strong is the mistaken link between decision theory and instrumental rationality that the erroneous idea that preferences are "tastes" (so rational agents have the goal of maximizing the satisfaction of their tastes) remains prominent even in sophisticated theorists, some of whom go so far as to talk about "tastes for fairness." Kaplow and Shavell (2002: 431) claim that "if individuals in fact have tastes for notions of fairness — that is if *they feel better* off when laws or events that they observe are in accord with what they consider to be fair — then analysis under welfare economics will take such tastes into account…" (Emphasis added). Apparently if satisfying one's preference for a fair outcome does not result in one's feeling better, welfare economics cannot take it into account.

[11] One interesting problem concerns whether the Better Chances axiom is inconsistent with some sorts of process-dependent moral criteria that require fair lotteries to distribute goods. See Diamond, 1976; Broome, 1991; Drier, 2004; Gaus, 2007. I think the

proper analysis of this type of problem is essentially the same as my analysis of the last-mango refuser.

[12] See note 3.

References

Anand, Paul (1993) *Foundations of Rational Choice under Risk* (Oxford: Oxford University Press).

Arrow, Kenneth J. (1982) Risk perception in psychology and economics, *Economic Inquiry* 20, pp. 1-9.

Benn, Stanley I. (1988) *A Theory of Freedom* (Cambridge: Cambridge University Press).

Benn and G.W. Mortimore (1976) Technical models of rational choice, in: S.I. Benn and G.W. Mortimore (Eds) *Rationality and the Social Sciences,* pp. 157-196 (London: Routledge and Kegan Paul).

Bolton, Gary E. (1991) A comparative model of bargaining: theory and evidence, *The American Economic Review* 81, pp. 1096-1136.

Broome, John (1991) Rationality and the sure-thing principle, in: Guy Meeks (Ed) *Thoughtful Economic Man*, pp. 74-102 (Cambridge: Cambridge University Press).

McMahon, Christopher (2001) *Collective Rationality and Collective Reasoning* (Cambridge University Press).

Clements, Paul and Emily Hauptmann (2002) The reasonable and the rational capacities in political analysis, *Politics & Society* 30 (March), pp. 85-111.

Craven, John (1992) *Social Choice* (Cambridge: Cambridge University Press).

Diamond, Peter A. (19967) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment, *Journal of Political Economy*, 75: 765-66.

Dreier, James (2004) Decision theory and morality, in: Alfred R. Mele and Piers Rawling (Eds) *The Oxford Handbook of Rationality*, pp. 156-181 (Oxford: Oxford University Press).

Gaus, Gerald F. (2002) Principles, goals and symbols: Nozick on practical rationality, in: David Schmidtz (Ed) *Robert Nozick*, pp. 105-130 (Cambridge: Cambridge University Press).

Gaus, Gerald F. (2003) Once more unto the breach, my dear friends, once more: McMahon's attempt to solve the paradox of the prisoner's dilemma, *Philosophical Studies* 116, pp. 159-170.

Gaus, Gerald F. (2007) *On Philosophy and Economics* (Belmont, CA: Wadsworth).

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

Hammond, Peter (1988) Consequentialist foundations for expected utility, *Theory and Decision* 25 (1988), pp. 25-78

Hampton, Jean E. (1998) *The Authority of Reason* (Cambridge: Cambridge University Press).

Kaplow, Louis and Steven Shavell (2002) *Fairness versus Welfare* (Cambridge, MA: Harvard University Press).

Luce, R. Duncan and Howard Raffia (1957) *Games and Decisions* (New York: John Wiley & Sons, 1957).

McClennen, Edward (1990) *Rationality and Dynamic Choice* (Cambridge: Cambridge University Press).

Morrow, James D. (1994) *Game Theory for Political Scientists* (Princeton: Princeton University Press).

Mueller, Dennis (2003) *Public Choice III* (Cambridge: Cambridge University Press).

Nozick, Robert (1993) *The Nature of Rationality* (Princeton University Press).

Pincione, Guido and Frenando Tesón (2001) Self-defeating symbolism in politics, *The Journal of Philosophy* 98 (December), pp. 636-652.

Rawls, John (1996). *Political Liberalism*, paperback edn. (New York: Columbia University Press).

Riker, William (1988). *Liberalism against Populism* (Prospects Heights, IL: Waveland).

Sen, Amartya (1970) *Collective Choice and Social Welfare* (San Francisco: Holden-Day).

Sen, Amartya (2002a) Introduction, in: Amartya Sen, *Rationality and Freedom,* pp. 3-64. (Cambridge, MA: Harvard University Press).

Sen, Amartya (2002b) Goals, commitment and identity, in: Amartya Sen *Rationality and Freedom*, pp. 207-224 (Cambridge, MA: Harvard University Press).

Sen, Amartya (2002c) Maximization and the act of choice, in: Amartya Sen *Rationality and Freedom*, pp. 159-205 (Cambridge, MA: Harvard University Press).

Skyrms, Brian (2004) *The Stag Hunt and the Evolution of Social Structure* (Cambridge: Cambridge University Press).

Tversky, Amos and Daniel Kahneman (2000) Rational choice and the framing of decisions, in: Daniel Kahneman and Amos Tversky (Eds) *Choices, Values and Frames*, pp. 209-223 (Cambridge: Cambridge University Press).