Rational Choice Theory

Brian Kogelmann and Gerald Gaus

1. Introduction

"The theory of justice is a part, perhaps the most significant part, of the theory of rational choice," writes John Rawls (1999: 15). In linking the theory of justice to rational choice Rawls both continued an intellectual tradition and began a new one. He continued an intellectual tradition in the sense that rational choice theory was first formally introduced in von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944), which launched a research agenda that continues to this day. Rawls, James M. Buchanan (1962; see Thrasher and Gaus, forthcoming) and John Harsanyi (1953; 1955) led the way in applying the idea of rational choice to the derivation of moral and political principles; a later generation of political theorists applied game theoretic analysis to a wide variety of problems of social interactions (e.g., Hampton, 1986; Sugden, 1986; Binmore, 2005). David Gauthier's *Morals by Agreement* (1986) was perhaps the most sustained and resolute attempt to apply models of utility maximization and decision theory to the derivation of social morality. Yet, despite the fact that some of the most respected political theorists of the last fifty years extensively employed the tools of rational choice theory, confusion still reigns about what these tools presuppose and how they are to be applied. Many reject the entire approach by a common refrain that "rational choice assumes that people are selfish," and since that is false, the entire line of analysis can be dismissed; a slightly more sophisticated (but still misplaced) dismissal insists that rational choice is about "preferences" but political philosophers are interested in "reasons," so again the entire approach can be set aside asserting that rational choice theory assumes selfishness, denying that people are selfish, and dismissing rational choice accordingly. A fundamental aim of this chapter is to explain just what the tools of rational choice theory are — and what presuppositions they make — and to provide some guidance for those wishing to follow in Rawls's footsteps, showing how rational choice theory can be applied to some of the fundamental issues of moral and political theory.

But even for those who do not want to explicitly engage in the contractarian project and thus follow in Rawls's footsteps, there are still important reasons to learn the basics of rational choice theory. First, one cannot properly engage with those who do explicitly rely on rational choice theory without understanding its basics. As just one example, arguing that parties in Rawls's original position would not choose the difference principle is fruitless unless one further engages with how Rawls defines the original position — given the tools of rational choice theory Rawls relies on, and given how the choice problem is defined, the *only* rational choice is the difference principle. Second, even if one does not want to primarily use rational choice theory in one's theorizing, knowledge of the theory can open up avenues of research to lend a supporting role. As an example, even though a deliberative democrat's core normative theory does not rely on rational choice in the way that Rawls's does, knowledge of the theory can help one engage, say, in the social scientific institutional design literature which could be relevant for the democrat's project.

In section 2 we explain preferences and utility functions as well as parametric and strategic choice. Section 3 sketches the dispute whether rational choice is an adequate mode of explanation, while section 4 highlights the way in which rational choice theory is normative theory, and how it has been employed in normative political philosophy. Section 5 seeks to draw our observations together in a checklist of decisions for those that would employ rational choice methods in their own work.

2. What is Rational Choice Theory?

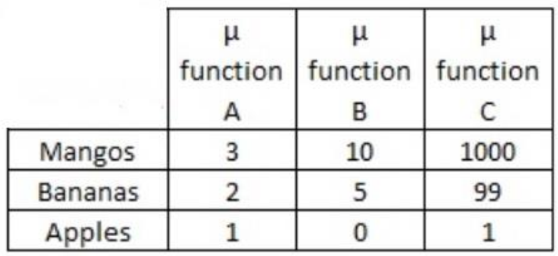
2.1 *Preferences and Utility Functions*

Rational choice is, at bottom, a theory of preference maximization. Many think of preferences as non-relational tastes or desires — having a preference for something is simply liking or desiring that thing. On this common view, to say that Alf has a preference for mangos means that Alf has a taste or desire for mangos; thus to say that he maximizes his preferences is just to say that he maximizes his pleasures or desire satisfaction. It is thus very common — for both defenders and critics — to see rational choice theory as inherently bound up with a Hobbesian or Benthamite psychology (Kliemt, 2009: 46ff; Sen, 2009: 178-83; Plamenatz, 1973: 20-7, 149-59). While such an understanding of preferences is perhaps colloquially natural, rational choice theory always understands preferences as comparative or relational, and they have no necessary connection to desires or pleasure. The core of rational choice theory is the primitive conception of a preference as a binary relation, of the form "x is preferred to y" ($x \succ y$). This is emphatically *not* a comparison of the strengths of two preferences, that for x and for y. It is literally incoherent in decision theory to claim one simply "prefers mangos" — a preference *is* a binary comparison.

Perhaps the best understanding of preferences is as deliberative rankings over states of affairs. When Alf deliberatively ranks (x), the state of affairs in which he eats a mango over (y), the state of affairs in which he eats a banana, we say that Alf prefers mangos to bananas ($x \succ y$). Note that we can leave entirely open the considerations Alf employed to arrive at this ranking: these could have been self-interest, desire-satisfaction, or his conception of virtue ("The brave eat mangoes, while only cowards eat bananas!"). Utility theory is a way to *represent consistent choice*, not the foundation of those choices. If Alf employs some deliberative criteria such that he ranks x above y for purposes of choice, then we can represent this as $x \succ y$; it is critical to realize that he does *not* rank x above y *because* he prefers x to y.

These preferences over states of affairs, combined with information, can be used to generate preferences over actions (say α and β). Thus in rational choice theory we can map preferences over "outcomes" (x, y) to preferences over actions (α, β), which, at least as a first approximation, can be understood as simply routes to states of affairs. Does Alf prefer grocery shopping at (α) Joseph's or (β) Caputo's? That depends on his information. If Alf knows that Joseph's has only mangos (x) and Caputo's has only bananas (y) then since $x \succ y$, $\alpha \succ \beta$ — he will prefer shopping at Joseph's to shopping at Caputo's.

From an individual’s preferences we can derive an *ordinal utility function*. Let us consider preferences over three fruits, mangoes (x), bananas (y) and apples (z). An ordinal utility function is a numerical representation of a person’s preferences, examples of which are illustrated in Table 1. We can derive an ordinal utility function for any individual so long as the individual’s preferences satisfy the following axioms.



	µ function A	µ function B	µ function C
Mangos	3	10	1000
Bananas	2	5	99
Apples	1	0	1

Table 1

1. Preferences are *complete*. For any two states of affairs x and y, either $x \succ y$, $y \succ x$ or, we can say, Alf is indifferent between x and y ($x \sim y$). We can define indifference in terms of what might be called the true primitive preference relation, “at least as good as.“ If Alf is indifferent between x and y ($x \sim y$) we can say that x is at least as good as y ($x \succsim y$) and y is at least as good as x ($y \succsim x$). We can also define “strict preference” ($x \succ y$) in terms of this more fundamental binary relation “at least as good as”: $x \succ y$ implies x$\succsim y$ and ¬[1]($y \succsim$x).

2. Strict preferences are *asymmetric*. $x \succ y$ implies ¬($y \succ x$). Indifference is, however, *symmetric*: ($x \sim y$) implies ($y \sim x$).

3. The true primitive preference is *reflexive*. Alf must hold that state of affairs x is at least as good as itself ($x \succsim x$)

4. Preferences must be *transitive*. If Alf prefers x to y and y to z, then Alf must prefer x to z ($x \succ y$ & $y \succ z$ implies $x \succ z$).

It is important to note that the numbers used to rank options in Alf’s ordinal utility function tell us very little. All an ordinal utility function implies is that higher-numbered

[1] Read “¬” as “not.”

states of affairs are more preferred than lower-numbered states of affairs. Turning back to Table 1, utility function A contains the same information as utility function B which contains the same information as utility function C. They all tell us that Alf prefers mangos to bananas to apples, nothing more. Thus ordinal utility information can be limiting. As we shall see, when modeling rational choice under risk and uncertainty we require information about (to put the matter rather roughly) the "distances between preferences." We need to know more than the fact that Alf prefers mangos to bananas — we must know (again, very roughly) how much more he prefers mangos to bananas. This brings us to *cardinal utility functions*, which contain such information. We can take any ordinal utility function and derive a cardinal utility function so long as preferences satisfy a few more axioms:[2]

5. *Continuity*. Assume again that for Alf mangos (x) are preferred to bananas (y), and bananas are preferred to apples (z), so $x \succ y$, $y \succ z$. There must exist some lottery where Alf has a p chance of winning his most preferred option (x) and a 1-p chance of receiving his least preferred option (z), such that Alf is indifferent between playing that lottery and receiving his middle option (y) for sure. So if such a lottery is L(x,z), we can say that for Alf [L(x,z)]$\sim y$.

6. *Better Prizes*. When Alf is confronted with two lotteries, L_1 and L_2, which (*i*) have the same probabilities over outcomes (for example, .8 chance of the prize in the first position, and so .2 chance of the prize in the second position), (*ii*) the second position has the same prize in both L_1 and L_2, and (*iii*) in the first position, L_1's prize is preferred by Alf to L_2's, Alf must prefer playing lottery L_1 to lottery L_2 ($L_1 \succ L_2$).

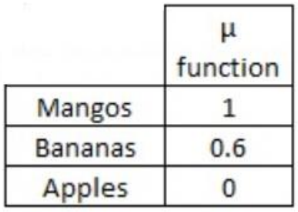
7. *Better Chances*. When Alf is confronted with two lotteries, L_1 and L_2, where (*i*) L_1 and L_2 have the same prizes in both positions (x and y, where $x \succ y$) and (*ii*) L_1 has a higher

[2] For simplicity sake, we state these in terms of the strict preference relation. A more general statement would use the "at least as good" relation. The intuitive ideas are clearer with strict preference.

chance (say .8) of winning *x* than L_2 (say .6), Alf must prefer playing lottery L_1 to lottery L_2 ($L_1 \succ L_2$).

8. *Reduction of Compound Lotteries*. Alf's preferences over compound lotteries (where the prize of a lottery is another lottery) must be reducible to a simple lottery between prizes. Thus the value of winning a lottery as a prize can be entirely reduced to the chances of receiving the prizes it offers — there is no additional preference simply for winning lottery tickets (say the thrill of lotteries).

If Alf's preferences obey the above axioms we can derive cardinal utility function in the following manner, originally proposed by von Neumann and Morgenstern (1944: chap. 3). Consider Alf's ordinal utility function, represented in Table 1. We take Alf's most preferred option, *x*, and assign it an arbitrary value, say 1. We then take Alf's least preferred option, *z*, and assign it an arbitrary value, say 0. We then take an option in-between (*y*) and compare Alf's preference between that option and a lottery between Alf's *x* and *z*. We manipulate the probability of the lottery until Alf is indifferent between that lottery and the middling option under consideration (the Continuity axiom guarantees there will be such lottery). Suppose Alf is indifferent between *y* and a lottery with *p* = .6 chance of winning *x* and, so, a 1-*p* (.4) chance of *z* (the probabilities in the lottery must sum to 1). Because the indifference obtains at this specific probability we assign a numerical value of .6 to the state of affairs in which Alf receives *y*, a banana. Alf's new cardinal utility function, representing how much more he prefers certain outcomes to others, is shown in Table 2.



	μ function
Mangos	1
Bananas	0.6
Apples	0

Table 2

Just how we are to understand this information is disputed. On a *very* strict interpretation, all we have found out is something about Alf's propensity to engage in certain sorts of risks; on a somewhat — but not terribly — looser interpretation we have found out something about the ratios of the differences between the utility of *x*, *y* and *z*, which can be further interpreted as information about the relative intensity of Alf's three preferences. It is important that our {1, .6, 0} utility scale is simply a representation as to how Alf views the relative choice worthiness of the options; he does not "seek" utility, much less to maximize it. He seeks mangoes, bananas and apples. To understand his actions as maximizing utility is to say that his consistent choices can be numerically represented so that they maximize a function. It is also critical to realize that the only information we have obtained is a representation of the ratios of the differences between the three options. There are an infinite number of utility functions that represent this information. Cardinal utility functions are thus only unique up to a linear transformation (which preserves this ratio information). If we call the utility function we have derived ***U***, then any alternative utility function ***U'***, where $\boldsymbol{U'} = a\boldsymbol{U} + b$ (where *a* is a positive real number and *b* is any real number), contains the same information. We can readily see why these utility functions are not interpersonally comparable: it makes little sense to simply add the utility functions of Alf and Betty, when each of their functions are equally well described in an infinite number of ways, with very different numbers.

2.2 *Parametric Choice*

After characterizing an agent's utility function we can go on to examine what sorts of choices rational agents make. The basic idea behind rational choice theory is that choosers *maximize expected utility*. Since utility, as we have seen, is simply a numerical representation of an agent's preferences, maximizing expected utility means that a rational agent maximizes the satisfaction of her preferences. Although this sounds straightforward — and it is, in certain contexts — maximizing one's preferences can be quite complex once we start examining choice under risk and uncertainty.

Let us first consider rational *parametric* choice. When an agent chooses parametrically it is assumed that the choices that the agent makes do not influence the parameters of other actors' choices (i.e., does not influence their preferences over outcomes) and vice versa. Their combined choices can affect her options (for example, the combined choices of other consumers determines prices), but when she acts she takes all this as fixed and beyond her control. As far as the chooser is concerned, she has a fixed preference ordering and confronts a set of outcomes mapped on to a set of actions correlated with those outcomes. This is opposed to *strategic* choice, examined in the next section. We can understand decision theory as the study of rational parametric choice, game theory the study of rational strategic choice.

In the simplest case of rational choice the individual is choosing under certainty. Here Betty not only knows her orderings of outcomes ($x \succ y \succ z$) — an assumption we make throughout — but also that she knows with certainty that, say, action α produces *x*, action β produces *y* and γ produces *z*. Given this she has no problem ordering her action-options $\alpha \succ \beta \succ \gamma$, and so as a rational utility maximizer she chooses α out of the set of options (α, β, γ) But we do not always choose under certainty. Sometimes we choose under *risk*. When choosing under risk the outcomes of our action options are not certain, but we do know the probability of different outcomes resulting from our choice act. Suppose Alf faces the option between action α (choosing covered fruit bowl A) and β (covered fruit bowl B). Bowl A might contain a mango or it might contain an apple – Alf does not know which. Bowl B might contain a mango or it might contain a banana. Once again Alf does not know which. Alf does know, however, that if bowl A is chosen then there is a .5 chance of getting a mango and a .5 chance of getting an apple. Alf also knows that if bowl B is chosen then there is a .25 chance of getting a mango and a .75 chance of getting a banana. Because Alf knows the probabilities that each action will produce different outcomes, Alf is choosing under risk. How does a rational person choose in risky situations?

In such situations rational agents use an *expected utility calculus*. With an expected utility calculus agents multiply the probability an action will produce an outcome by the utility one assigns to that outcome. We assume that Alf knows all the possible outcomes that each action option will produce. Suppose Alf assigns von Neumann and Morgenstern cardinal functions over outcomes as depicted in Table 2. With this above cardinal utility function Alf's expected utility calculus goes like this:

α: choosing bowl A: .5(1) + .5(0) = .5

β: choosing bowl B: .25(1) + .75(.6) = .7

Since the expected utility of α (choosing Bowl A) is .5 and since the expected utility of β (choosing) bowl B is .7 Alf chooses what is the best prospect for maximizing his preferences, which is β. Notice that this does not ensure that, in the end, he will have achieved the highest level of preference satisfaction: if he did choose α *and* it turned out to have the mango while β produced the banana, α would have given him the best outcome. The point of expected utility is that, at the time of choice, β offers the *best prospects* for satisfying his preferences.

It is sometimes thought that any set of cardinal numbers are sufficient for expected utility calculations, as we can multiply cardinal numbers by probabilities (you can't multiply ordinal utilities). This is too simple. Our von Neumann–Morgenstern cardinal utility functions possess *the expected utility property*, as they take account of Alf's attitude towards risk. Recall that von Neumann–Morgenstern functions are generated through preferences over lotteries, thus they include information about how Alf weighs the attractiveness of an outcome and the risk of failing to achieve it. Suppose that instead of employing the von Neumann–Morgenstern procedure we asked Alf to rate on a scale of 0 to 1, simply how much he liked mangoes, bananas and apples, and he reports 1, .6. and 0 (this *looks* just like Table 2). If we then multiplied, say, this type of "cardinal utility" of 1 for mangoes times a .5 probability that α will produce a mango, we would get the *expected value* of α, but not its expected utility, for Alf could be very reluctant to take a .5 chance of his worst outcome (an apple) in order to get a .5 chance of his best; thus so to him the utility of α could be far less than .5. But our von Neumann-Morgenstern procedure already factored in this information. Only in the special case in which people are *risk-neutral* (they are neither risk prone nor risk averse) will expected value be equivalent to expected utility.

Choice under *uncertainty* is a messier and more complex affair. Like choice under risk, when choosing under uncertainty one does not know the outcome of one's act of choice. But unlike choice under risk, one also does not know the probabilities that an action-option will yield the various possible outcomes. Following Luce and Raiffa (1957: 278) we note that the idea of uncertainty is vague. With choice under risk we know with objective certainty the relevant probabilities that α will produce various outcomes (.5 of *x,* .5 of *z*). What happens, though, if we are not certain about the relevant probabilities but more-or-less have a reasonable guess as to what the probabilities are? Is this choice under uncertainty? Technically speaking it is. Instead of having two categories, risk and uncertainty, it is more helpful to think of uncertainty as coming in degrees. Though we do not know the probabilities of outcomes occurring, our credence in what the relevant probabilities are can be more or less founded. At one end of the spectrum is choice under risk, where we know the probabilities of any given action option (so that this sums to 1 for each act), and on the other end of the spectrum is choice under radical uncertainty, where we cannot even begin to intelligently guess the relevant probabilities over outcomes.

Most relevant for political theory is choice under radical uncertainty. As we shall see later on, one of the central debates in twentieth century political theory largely revolved around what rational choice requires when one faces such uncertainty. Suppose, as before, that Alf must choose between α and β (on the basis of their associated fruits bowls). But unlike before, Alf has absolutely no basis for assigning probabilities as to which fruit the respective bowls might contain. In such a situation how does Alf choose rationally? It is here that rational choice theory turns more into an art than a deductive system. Although it is relatively straightforward how rational persons choose under risk, and somewhat less certain under uncertainty, it is by no means obvious how rational agents should choose when faced with radical uncertainty. Consider two different ways in which one might approach such choices, outlining the costs associated with the different methods.

(*i*) When facing radical uncertainty one might simply try to extend the choice procedure used in decision under risk — that is, one might employ an expected utility calculus. But an expected utility calculus requires that we assign probabilities to outcomes and, by hypothesis, when facing radical uncertainty there is no educated basis for assigning such probabilities. In such cases the *Laplacean principle of indifference* says to assign equal probability to all possible outcomes occurring, reflecting indifference to these outcomes occurring. Continuing our example, using this Laplacean principle Alf would assign an equal probability of α yielding a mango and an apple, and to β an equal probability of it giving him a mango and a banana. From there Alf performs his expected utility calculus as he did with choice under risk, so $\alpha = .5(1) + .5(0)$; $\beta= .5(1) + .5(.6)$, again yielding the choice of β.

The Laplacean principle of indifference is controversial. Ken Binmore (2009: 129) points out the ambiguity present in assigning equal probability to different outcomes occurring. Suppose Alf attends a three horse race between α, β and γ, but has no basis for assigning probability as to which horse will win. Is the Laplacean principle of indifference to be applied to the question (*a*) "which horse will win?" or (*b*) "will α win or not win?," "will β win or not win?", "will γ win or not win?" If the first, the Laplacean principle says Alf should assign each horse a ⅓ probability of winning (so that Alf is indifferent between which horse wins). If the second, the principle say to assign each horse a ½ probability of winning and ½ probability of not winning (so that Alf is indifferent between each horse winning and not winning).

(*ii*) Another way of choosing under radical uncertainty is by using the maximin decision rule. The maximin decision rule eschews the use of probabilities; it instructs one to examine each action option (α, β, γ) and identify for each its worst possible outcome (say α_W, β_W and γ_W). Looking only at these "minimum payoffs" for each action one is to select the highest among *them* (hence one should "maximize the minimum"). Like the Laplacean principle of indifference there are many criticisms of the maximin decision rule, most of which offer examples in which its application seems irrational. Suppose, following Rawls (1999: 136), one is faced with a choice between lottery L_1 and lottery L_2 whose probabilities are unknown. Suppose further the payoffs for L_1 and L_2 are in dollars (and that preferences are monotonic over money). This decision problem is shown in Figure 1:

	Prize 1	**Prize 2**
L_1	0	n
L_2	$1/n$	1

Figure 1: Rawls's Problem for Maximin

Maximin says to choose L_2 because L_2 guarantees a minimum payoff of $1/n$ while the minimum payoff for L_1 is 0. And when n is small (say, \$1, \$2, or \$3) this does not seem to be an unreasonable choice, but suppose n is quite large, perhaps \$1,000,000. In this case there does seem something irrational about choosing L_2 and thus something irrational about the maximin decision rule in general.

2.3 *Strategic Choice*

When choosing the decisions of others are often a fixed background; one does not have to include in one's deliberations the different things they might do. In such cases of parametric choice maximizing one's preferences is relatively straightforward, at least until we get to choice under uncertainty. But many times the actions of agents affect each other's decisions. This is strategic choice, which is studied by *game theory*. Roughly, a game is defined by (*i*) a set of players, (*ii*) a set of strategies for each player, (*iii*) the information available to the players and (*iv*) a set of payoffs for each player which is defined by her preference ordering (ordinal or cardinal) and which are mapped on to her strategies. Change any of these and you change the game.

Game theorists originally focused on zero-sum games, interactions of *pure conflict*. Whatever one player wins the other player loses. To see this consider Figure 2, which is the classic game Matching Pennies.

		Betty	
		Heads α	*Tails* β
Alf	*Heads* α	1 –1	–1 1
	Tails β	–1 1	1 –1

Figure 2: Matching Pennies

By convention, Row's (Alf's) payoffs are in the bottom left of each cell, Column's (Betty's) in the upper right. Note that the gains and losses in every cell equal 0; any gain must come from the other player, so we have a *zero*-sum game. John von Neumann (1928) proved that all zero-sum games with finite strategies have equilibrium solutions. That is, there is a rational course of action in all such games. Relating back to decision under uncertainty in parametric choice, the equilibrium solution for zero-sum games is to play one's maximin strategy – *to maximize one's minimum.* If both players maximize their minimum then neither player can gain from unilaterally changing her strategy. If we accept that the equilibrium solution concept models rationality then, when playing zero-sum games, it is rational to maximin.

Though the theory of games was first developed in zero-sum form it was quickly realized that zero-sum games are of limited applicability for understanding most strategic interactions. Rarely are we locked into total conflict; we often find ourselves in situations where we are competing *and* cooperating. Consider for example Luce and Raiffa's (1957: 90-1) politically incorrect "Battle of the Sexes" game. In this game Alf and Betty wish to go out with each other on a Friday evening; the worst thing for each would be to not go out together. However, Alf would prefer a date at the prize fights, Betty at the ballet. Figure 3 gives an example of the game, using cardinal utility.

		Betty	
		Go to Fights α	*Go to Ballet* β
Alf	*Go to Fights* α	1 2	0 0
	Go to Ballet β	0 0	1 2

Figure 3: the "Battle of the Sexes" in Cardinal Utility

So we have two players Alf and Betty, and each has the same two "pure" strategies, {Alf α, Alf β}, {Betty α, Betty β}. We suppose that each knows the information set **S**: they know each other's utilities and what strategies are open to the other, each has knowledge of the other's rationality, and they know that each must choose without knowing what the other has chosen. Moreover they each know **S′**: each knows that the other knows all the facts in **S**. And indeed they know **S″**: each knows that the others know all the facts in **S**′. The *common knowledge assumption* is assumed to iterative — we can go on with higher and higher levels, where each knows what the facts were at the previous level.

If Alf goes to the fights, then Betty, given her own preferences, will do best by also going to the fights; if Betty goes to the ballet, Alf will do best by also going to the ballet. Neither knows what the other is doing, but they wish to coordinate. So this is a game of cooperation. Yet it is also a game of conflict: they disagree about the best outcome. The central concept in game theory is that of *Nash equilibrium*. We can think of the Nash equilibrium as the solution to a game insofar as when two players are playing their equilibrium strategies each has made his/her best response to the move of the other player, and so neither player has any incentive to unilaterally change his/her strategy. The strategic interaction, we might say, can come to rest there. In the Battle of the Sexes Game there are *two* Nash equilibria in pure strategies: both play α or both play β. Players can also play a "mixed strategy," which is a probability distribution over their pure strategies. In Figure 2's game this would mean Alf would play α with a *p* probability and β with a 1-*p* probability; this game has a Nash equilibrium in mixed strategies in which Alf plays α with a probability of ⅔ (and so β with a probability of ⅓) and Betty plays β with a probability of ⅔. At this point no adjustment to the probability that they play each of their pure strategies will allow a player to unilaterally improve her expected payoffs. John Nash (1950a, 1951), in a fundamental theorem of game theory, showed that all games in cardinal utility with finite strategies have at least one equilibrium (sometimes only a mixed equilibrium). Note that this theorem, because it requires mixed strategies, depends on the expected utility property.

We should note a more general solution concept, *correlated equilibrium* (Aumann, 1987), which includes the Nash solution. Its attractions are manifest in the game of chicken:

		Chevy	
		Swerve α	*Don't Swerve* β
Ford	*Swerve* α	5 5	10 0
	Don't Swerve β	0 10	–5 –5

Figure 4: Chicken in Cardinal Utility

Suppose that every Saturday and Sunday two teenagers, one driving a Chevy and the other a Ford, drive toward each other with the pedal to the metal, and the first one to swerve is "chicken." So the options are between swerving (α) and not swerving (β). The best result is to keep driving straight while the other swerves: the other chickened, you didn't. If both swerve, neither crashes; if neither swerve, they crash. In this game the two "pure equilibria" are {Ford α, Chevy β} and {Ford β; Chevy α}. There is a mixed Nash equilibrium where each swerves half the time, but this has the unfortunate result that ¼ of the time they crash! Now consider a game where we add a strategy: on Saturdays Chevy swerves, on Sundays Ford swerves. The game will not be terribly exciting but they will never crash, and each will have an expected payoff of 5, as good as one could do against a rational opponent. But note that we seemed to have changed the game: we have expanded the strategy set to include a correlating device (the day and the make of car).

At least in their simpler forms, Chicken and the Battle of the Sexes raise problems when we understand strategic rational choice as requiring agents to play their equilibrium strategies. For in these games there are *multiple* equilibria. What, then, do rational players do? Which equilibrium do they play? Unfortunately there is no definitive good answer. People have tried many ways to select one equilibrium from multiple Nash equilibria. The oldest way of solving the equilibrium selection problem is to *a priori* theorize about what rationality requires when faced with multiple equilibria. This is known as "equilibrium refinement" and was pursued to its fullest by John Harsanyi and Reinhard Selten (1988). Some refinements seem plausible: for instance, if one equilibrium strictly dominates (is better for *all* parties) another, then perhaps that equilibrium is more rational. At best, this is a rather limited refinement. As we proceed we encounter deep controversy — no satisfactory account has been given to refine all games down to a unique equilibrium.

Others have employed evolutionary accounts explaining why some equilibria are selected over others: though a pair of strategies might be in equilibrium for a one-shot game, such a pair of strategies might no longer be in equilibrium if the game is continually repeated (the opposite is also true). Notable here is Brian Skyrms's (1996; 2004) work. In contrast, Thomas Schelling (1960: 57) emphasized the role of *salience* or the sheer obviousness of a solution in certain contexts. For example, when two parties are asked to split up a sum of money they will often split it right down the middle even though there are many possible splits, or if two parties are asked to meet at some place in Paris, they will usually go to the Eiffel Tower because it is such a large, obvious landmark. Of course, there can be multiple salient equilibria. Moreover, differing methods of equilibrium selection can conflict. Strict dominance as a plausible refinement technique can conflict with Schelling's salience — sometimes the salient equilibrium will be dominated by a non-salient equilibrium. Though there have been many proposals of how to select among multiple equilibria, a current limitation in the theory of games is that it cannot tell us which equilibrium rational players select.

One final area of strategic choice relevant for political theory is bargaining games. Implicitly we have been examining *non-cooperative* game theory, which assumes that choices among strategies are not binding. One follows a Nash equilibrium not because it is the solution to a game, but because, given the move of the other, it is one's best available move. In *cooperative* game theory, choices among strategies are assumed to be binding by some external enforcement mechanism. The most famous case of a cooperative solution that is not in equilibrium is the Prisoner's Dilemma. The story is normally told in terms of prisoners getting caught and seeking to minimize jail time, but all games are really about utility. Figure 5 gives the Prisoner's Dilemma in terms of cardinal utility.

		Betty	
		Don't Confess α	*Confess* β
Alf	*Don't Confess* α	x x	1 0
	Confess β	0 1	y y

Where $1>x>y>0$

Figure 5: The General Form of the Prisoner's Dilemma in Cardinal Utility

If Betty plays α, Alf will get *x* if he plays α, and 1 if he plays β; since 1>*x*, he should then play β. If Betty plays β, he gets 0 if he plays α, and *y* if he plays β; since *y*>0, he should again play β. "Confessing" (β) is thus his a *dominant strategy*: no matter what Betty does, Alf does best if he plays β. And it can clearly be seen that Betty will reason in exactly the same way. So {β,β} is the sole *Nash equilibrium.* Yet the outcome of *y/y* is worse for each than the *x/x* outcome. The *x/x* is a cooperative outcome that is not in equilibrium; as a noncooperative game there is only one solution to the Prisoner's Dilemma: *confess!* Such games can, however, form the basis of binding agreements that yield a cooperative surplus. Faced with such bargains one might ask what distribution of surplus rational players will settle on. The status of bargaining theory is unsettled. There are several competing bargaining theories. Nash (1950b)'s bargaining solution has the widest support. And Ariel Rubenstein (1982) appeared to put cooperative bargaining theory on a firmer footing by clearly showing how it can be derived from noncooperative game theory. Yet Luce and Raiffa's (1957: 119-120) worry remains — that such bargaining solutions, though intellectually interesting, do not seem to model what purely rational players would select in actual bargains. And some worry that the very axioms of bargaining theory go beyond rationality to include some criteria of fairness (Thrasher, 2014).

3. Is Rational Choice Theory Unrealistic?

3.1 *People Aren't Always Rational*

It should be manifest that rational choice theory idealizes: agents are generally understood to know all their options, rank them, and usually in a way that provides a cardinal scale. They are then assumed to be thoroughly consistent choosers, always seeking to maximize the satisfaction of their preferences. The most common refrain is that, since people are not thoroughly rational, these are wildly unrealistic assumptions, and so undermine the usefulness of rational choice theory. If people aren't rational, why care about rational choice?

One version of this common refrain simply points out that people are not *always* rational: they do not always act as rational choice theory predicts. First, we need to be clear about what rational choice theory does *not* predict: it does not predict that people will always be self-interested, or that they seek to maximize the satisfaction of their desires, or that moral considerations never move them. These are matters about the nature of the deliberation that determine a person's rankings — they concern the basis on which Alf determined whether state of affairs *x* is to be ranked above *y.* As we have stressed, rational choice theory is not committed to any account about the basis of rankings. Recall also that preferences are not desires, but rankings of outcomes, which determine preferences over actions. This is important. Many experimenters, for example, believe that they have shown that people cooperate in Prisoner's Dilemmas. In experiments in which people are confronted with situations that *mimic* Figure 5, where the payoffs are monetary and players are awarded sums of money where the size of the monetary payoffs is exactly correlated with the utility payoff schedule in Figure 5, we can observe people cooperating — an action that, for all players, is dominated by confessing. However, unless the players rank the outcomes *only* in terms of their monetary payoffs — unless their preferences are *only* over amounts of money, and thus for example have no concern with fairness or simply not being a jerk — we cannot say that the players are actually in a Prisoner's Dilemma interaction, so rational choice theory does not predict that they will defect.

Nevertheless, even when we are more careful, there is dispute over whether rational choice theory is a good predictor. Perhaps the most comprehensive critique of rational choice theory as a predictive tool is Ian Shapiro and Donald Green (1996)'s *Pathologies of Rational Choice Theory* (see also Amadae, 2015), though there have been equally comprehensive defenses of rational choice theory as a predictive tool (see, for instance, the debate in Friedman, 1996). Rational choice theory is a *model* of human action: like all models, we build a simplified world so that we can better understand some salient dynamics, and get some predictive leverage in some contexts. As Michael Wesiberg (2007) notes: "[w]hen faced with…complexity, theorists can employ one of several strategies: They can try to include as much complexity as possible in their theoretical representations. They can make strategic decisions about which aspects of a phenomenon can be legitimately excluded from a representation. Or they can model, studying a complex phenomenon in the real world by first constructing and then studying a model of the phenomenon." These models abstract from the complexity of reality to analyze the interactions among key elements. In some contexts the models may abstract from too much of the complexity to give adequate understanding of the phenomena, but in other contexts they can be enlightening and accurate. In social science no model or theory explains most of the variance of interestingly complex phenomena, but this does not mean the model is without predictive and explanatory power. Rational choice models do rather well in a variety of cases, especially where we can make accurate suppositions as to what people's utility functions look like, the information sets they have, and the actions open to them. Thus experiments have conformed that various sorts of auction behavior and trading behavior conform to the predictions of rational choice theory (Plott: 1986; Smith: 2000). On the other hand, when people have complex and heterogeneous utility functions, when the options are not well understood or, more radically when they act out of strong emotions, rational choice theory is not apt to explain things well (Schwartz: 2002; Bosman, Sutter and van Winden: 2005).

3.2 Rationality and Intelligibility

A more radical rejection of rational choice theory insists that rationality is a "folk concept," not a scientific one. While most of us, as normal "folk", think in terms of preferences, options and choices, some argue a "true" scientific theory would be purely causal, and so would explain human actions in terms of psychological or biochemical processes. Hartmut Kleimt advances what might be understood as a "compatibilist" reply: "humans are citizens of two worlds, a lower 'material' world and a higher one of reason" (2009: 16). On this compatibilist approach, we can understand humans both in simply causal terms and in purposive ways — as agents with preferences — who seek to secure their most favored outcome. Some suggest that we only rely on rational choice models because our causal models are underdeveloped, so perhaps someday we can do without rational choice theory. But this seems wrong. Humans understand each other through "mind reading": the actions of another are not really intelligible to me until I see how, if *I* thought what she believes and *I* cared for what she cared for, *I* would have done what she did. Making others intelligible to us is closely bound to seeing them as rational: rational action makes sense to us (Rosenberg, 1995). True, sometimes it is intelligible to us why people are not rational: we can understand all too well why someone who is drunk accepts a dangerous and silly dare. But usually, when we are confronted by simply irrational behavior we do not understand what it is really all about. To understand the behavior of another is to see it as intelligible, and to see it is rational is to render it intelligible.

4. Rational Choice as a Normative Theory

4.1 *Deliberative Rationality*

Herbert Gintis (2009: chaps 1, 12) boldly conjectures that game theory's axioms of consistent choice express, at the deepest level, the ingredients of evolutionary success. It is indeed striking how powerful, in the guise of "evolutionary game theory," the tools of game theory have proven in modeling natural selection (for classic statements, see Smith, 1975; Hamilton, 1964; Trivers, 1971). If the basic idea of consistent choosers who maximize their satisfaction of their goals is foundational to selected behavior in humans and non-humans alike, we might hypothesize that creatures like us — who are conscious of their agency — cannot help but see these axioms as normative: as not simply explaining what they do, but guiding what they *should* do. In a secular world where the gods no longer seem to speak to, and guide us, can we have any deeper guide than our most basic axioms of successful agency, upon which perhaps our entire evolutionary past hinges? Thus is the crux of Gauthier (1991)'s case for what he calls "contractarianism": the only truly solid foundation for thinking about what we ought to do that our naturalistic and scientific age has left us are the consistency axioms at the root of successful agency. If we have any hope of identifying normative principles for our post-metaphysical world — a world shorn of faith in God's plan for us, or invisible normative properties that instruct us what to do — they must be solidly grounded in rational choice.

Contractarianism in moral and political theory seeks to show that moral and political principles or rules can be the result of an agreement between rational parties and, further, that compliance with these principles or rules is rational. When deriving normative principles contractarians appeal to what rational parties would agree to, which is just to say what rational choice requires in strategic interaction. This can be seen informally in the case of Hobbes when he argues that rational parties, engaged in a war of all against all in the state of nature, would agree to enter the social contract, where each gives up her right to all things in exchange for security from the sovereign (1651: chaps. 13, 16). Contemporary contractarians go further than examining the informal logic of what rational parties would agree to. Gauthier (1986), for instance, employs game theoretic bargaining models to examine what moral principles rational parties would agree on. He ends up developing his own bargaining solution, the principle of minimax relative concession, which is an offshoot of the Kalai-Smorodinksy (1975) bargaining solution. Most contractarians — including Gauthier at a point in his career — opt for the alternative, Nash bargaining solution (1950b). But the critical idea was that this bargain would simply be a rational bargain: the only normativity is the normativity of rationality.

Just as we had to define Alf's utility function before examining what a rational Alf would choose, contractarian approaches to moral and political theory also require that we give content to utility functions. Before showing that rationality leads to certain normative principles we must have some account of what people's preferences are. Hobbes (1651: chap. 11) does this when he posits in all mankind "a perpetual and restless desire for power after power, that ceaseth only in death." Gauthier (1986: 87) also does this when he insists that utility functions be non-tuistic, or not other-regarding. This means that the satisfaction of one's preferences cannot rely on other people's preferences being satisfied. Excluded by non-tuism is the mother who only prefers to see her children's preferences satisfied. The reason why Gauthier excludes such preferences is an attempt to avoid his theory admitting solutions based on adaptive preference formation under domination: for example, Sen (2009: 166-67, 285-6) notes how in some areas Indian women's understandings of what they want and how well they are doing are fundamentally shaped by their acceptance of the normality of gender bias.

Contractarians such as Hobbes and Gauthier are the purest examples of the project of building moral and political principles on the normativity of rational choice. Rawls's famous social contract theory also seeks to base the normative appeal of their conclusions on the normativity of rational choice. At the outset of *A Theory of Justice* Rawls (1999: 16) insists that "one conception of justice is more reasonable than another, or justifiable with respect to it, if rational persons in the initial situation would choose its principles over those of the other for the role of justice. Conceptions of justice are to be ranked by the acceptability to persons so circumstanced. Understood in this way the question of justification is settled by working out a problem of deliberation: we have to ascertain which principles it would be rational to adopt given the contractual situation. This connects the theory of justice with the theory of rational choice." And thus the importance of our opening quote from Rawls (1999: 15): "the theory of justice is part, perhaps the most significant part, of the theory of rational choice."

Rawls, however, insists that the choice of moral principles is not simply governed by the "rational" but also the "reasonable" — a concern with fair cooperation (1993: 48ff). One way to model this is to select principles behind a veil of ignorance, in which parties have drastically restricted information sets, including information as to who they are, and what their utility functions look like (Gaus and Thrasher, 2015; Harsanyi, 1955). This restriction on information forces parties to choose fairly: hence non-tuistic parties, i.e., choosers only concerned with maximizing their own goals, are forced (via ignorance) to consider how the goals of *each* fare, since once the veil is lifted, one may end up being any of these people. A rather more straightforward route, perhaps, is simply to build concerns with fairness or morality into the utility functions of the choosers (Gaus 2011: chap. 5). Again, it is crucial to stress that rational choice theory does not tell us what a person's preferences are. The more one stresses the "reasonable" over the "rational," the more one is insisting that the selection of principles must not only suppose utility maximizers, but those with certain substantive moral concerns.

5. Decisions to be Made When Using Decision Theory

We have tried to stress the complexities of rational choice theory, and have tried to alert readers to the pitfalls of widespread misconceptions. We can apply these lessons to thinking about how a researcher might contemplate using a rational choice framework in her own work in political theory.

- It may sound terribly simple, but **the first step is to be clear whether you are using rational choice theory as an explanatory, predictive model or as a normative model**. These can be surprisingly easy to run together. Think of Hobbesian political theory: part of Hobbes's famous argument is how people will act in the state of nature (an explanatory project) and the next part is a normative contract: what would be a rational agreement to leave the state of nature? These raise vastly different issues. The explanatory part of the project has been advanced by game theoretic modeling concerning the ease with which conflict can erupt in a state of nature (Vanderschraaf, 2006; Chung, forthcoming). The specification of the contract is normative: what constitutes a rational bargain? This leads to the sorts of concerns Gauthier and Rawls have investigated. Confusion between these two projects can lead errors. In offering an empirical analysis of the state of nature Hobbesians are not saying that this is what persons *ought* to do – in fact, the Hobbesian project is driven by a desire to show how alleviating one's self from such a state is consistent with rational choice. And in offering normative analyses of the social contract theorists are not trying describe anything that has, or necessary will, happen. Instead, they are trying to model the reasons we have to obey certain authority relations via a rational choice analysis.

- **Be clear about whether the problem is parametric or strategic.** Even the best political theorists can make this mistake. Rawls (1999: 11) said that we should understand the derivation of the theory of justice as the result of a "fair agreement or bargain," suggesting that Rawls wanted to model a strategic rather than parametric choice problem in the original position. But later on Rawls (1999: 120) insists that, due to the normalization of the parties' interests, we can *really* understand choice in the original position "from the standpoint of one person selected at random," suggesting a parametric rather than strategic choice problem. This can lead to confusion. Consider, for instance, Rawls (1999: 132-133)'s continual insistence that we can think of the two principles of justice as "those a person would choose for the design of a society in which his enemy is to assign him his place." If Rawls models the original position as a parametric choice problem then such remarks conjure up the maximin choice principle under uncertainty – which, as we have seen, is riddled with controversy. But suppose Rawls models the original position as a strategic choice problem rather than a parametric one. Then, if the game is zero-sum, we have seen that the maximin choice *is the uniquely rational solution to the game*, and thus not subject to controversy. Whether the original position is parametric rather than strategic is thus highly relevant in terms of its plausibility as a rational choice model. We end up following the standard way of understanding the original position in the literature as a parametric choice problem rather than a strategic choice problem though, again, Rawls's language could be clearer on this point.

- **The really hard step is to determine the utility functions of the parties.** And how utility functions are defined can have tremendous relevance for the plausibility of a rational choice model, regardless of whether the model is normative or explanatory. Consider again Gauthier's normative contractarian project. As we mentioned, Gauthier defines utility functions by requiring preferences be non-tuistic so that his bargain is not the result of preference adaptation under domination. Such stipulation, though, is not without controversy. Hubin (1991) argues that excluding tuistic preferences means that rational persons with such preferences will not find it rational to comply with the resulting bargain. After all, if the bargain was derived from a non-tuistic utility function, but a rational person has a tuistic utility function, then it is hard to see why it is rational for the other-regarding person to obey constraints derived specifically for persons who are not other-regarding. And from the explanatory side, Hampton (1986: 75) argues that, given one interpretation of how Hobbes defines utility functions of actors in the state of nature, there would actually be no conflict in the state of nature at all. If people really were rational actors who only cared only about self-preservation, then the iterated context of the state of nature would allow for cooperation. No need for the sovereign after all!

- **You should now decide whether to represent these preferences ordinarily or cardinally.** For some simple games (such as the Prisoner's Dilemma) an ordinal representation will be sufficient; but if risk or uncertainty is involved, we have seen, a cardinal representation is likely to be required.

- **Your next decision is: what are the actor's information sets? Do they know everything, or only a little?** And if one is modeling strategic interaction, one must decide if what the parties know is the same, or if some parties know more than others. This latter difference can be crucial, again for both normative and explanatory rational choice models. On the normative side we again turn to contractarianism. One common assumption in bargaining theory is the assumption that parties are "symmetric" – that they have the same strategies available to them, same bargaining ability and, importantly, the same information sets. This assumption was originally introduced by Nash (1950b), but is common in many bargaining models, including Gauthier (1986). Yet as Thrasher (2014) shows, this assumption can often be at odds with the contractarian project – it amounts to importing fairness criteria into the terms of the bargain, when really what the contractarian is trying to do is *derive* fairness criteria *from* the bargain itself. For the reason why the symmetry assumption is often introduced is because once we admit things like asymmetric information sets into the bargain, the resulting distribution might not be so intuitively attractive when compared to symmetrical case. And from an explanatory standpoint, Ernst (2005) forcefully argues that many explanatory models of the evolution of conventions and fairness norms obscure more than they reveal when the assume symmetry across parties, of which symmetrical information sets is one such part. These sorts of considerations are highly relevant when determining the information sets of the parties – symmetrical information might be clean and easy, but possibly problematic.

- **If you decide that you are concerned with a strategic choice, you will need to learn a bit of game theory.** There are a number of good introductory texts (e.g., Maschler, Solan, and Zamir, 2013). **Do not be too quick to focus on a particular game before you have determined the preferences and the information sets of the parties.** Games are defined by the preferences, information sets and strategies of the players. It is all-too-common for even post-graduate students (and, alas, professors!) to get fascinated by a game such as the Prisoner's Dilemma or Chicken, and then to try to find ways to apply it, often forcing a case or situation into an inappropriate game. Once one has made the previous decisions on our list one can fruitfully think about game theoretic representations.

A last lesson: **when using rational choice theory, especially in normative contexts, it is critical to realize its limits.** Often our disputes are not about what a rational person would do, but about the correct preferences to ascribe to people, or the conditions for fair bargains. Understanding moral and political principles as an object of choice among rational choosers can clarify their basis, but we must be very careful not to reduce all these to disputes about rationality. Even experts can believe that their dispute is about rationality, rather than the proper content of an agent's concerns. A good example of this is the Rawls-Harsanyi debate. Harsanyi (1953, 1955) and Rawls pursued very similar contract-based normative projects, reducing the problem of moral justification to one of rational choice behind the veil of ignorance. *Contra* Rawls, however, Harsanyi believed that rational parties choosing behind a veil of ignorance would select average utilitarianism rather than Rawls's two principles of justice. At first blush the dispute between Rawls and Harsanyi reduced to an argument about rational choice under uncertainty. Rawls thought rational parties, when faced with radical uncertainty, ought to adopt the maximin decision rule. Harsanyi thought rational parties, when faced with radical uncertainty, ought to adopt the Laplacean principle of indifference. The debate continued after both parties published their seminal social contract contributions. Harsanyi (1974) continued to point out the irrationality of the maximin decision rule, while Rawls (1974) continued defending it.

Recent analysis of the Rawls-Harsanyi dispute, though, suggests that the disagreement between the two cannot merely be reduced to a debate about rational choice. Moehler (2013) argues that Rawls's and Harsanyi's disagreement over what rationality requires *itself* reduces to a philosophical dispute about which moral values should be modeled in the choice situation, and thus which moral values are important. Rawls wanted to model *impartiality* in the original position. Harsanyi was especially concerned with modeling *impersonality*. While impartiality merely requires that agents do not unjustifiably favor their own interests in making decisions, impersonality requires agents to make decisions only on the basis of the common interest. This inclusion of impersonality, which Rawls (1999: 24) rejected, *required* Harsanyi to use the Laplacean principle of indifference. And as Rawls (2001: 97-100) notes later on, the use of maximin reasoning in the original position ultimately comes down to a moral, not a rational, choice. The putative debate between

Rawls and Harsanyi over the nature of rational choice theory reduces to a disagreement over what is to be valued – impartiality versus impersonality. This is an object lesson. **Rational choice theory by no means replaces philosophical analysis. Remember always that it is a tool, and like all tools, is effective only when it is used to perform its proper tasks in its proper ways.**

Works Cited

Amadae, S. M. (2015). *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. Cambridge: Cambridge University Press.

Aumann, Robert. (1987). "Correlated Equilibrium as an Expression of Bayesian Rationality." *Econometrica 55: 1-18.*

Binmore, Ken. (2005) Natural Justice. Oxford: Oxford University Press.

—— (2009). *Rational Decisions*. Princeton: Princeton University Press.

Bosman, Ronald Matthias Sutter and Frans van Winden (2005). "The Impact of Real Effort and Emotions in the Power-To-Take Game. " *Journal of Economic Psychology,* vol. 26: 407–429.

Buchanan, James M. and Gordon Tullock. (1962 [2004]). *The Calculus of Consent: The Logical Foundations of Constitutional Democracy*. Indianapolis: Liberty Fund.

Chung, Hun. (forthcoming). "Hobbes's State of Nature: A Modern Bayesian Game-Theoretic Analysis." *Journal of the American Philosophical Association*.

Ernst, Zachary. (2005). "A Plea for Asymmetric Games." *Journal of Philosophy* 102: 109-125.

Friedman, Jeffrey, ed. (1996). *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. New Haven: Yale University Press.

Gaus, Gerald (2011). *The Order of Public Reason*. Cambrodge: Cambridge U uversity Press.

Gaus, Gerald and John Thrasher. (2015). "Rational Choice and the Original Position: The (Many) Models of Rawls and Harsanyi." In *The Original Position*, edited by Timothy Hinton. Cambridge: Harvard University Press.

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

—— (1991). "Why Contractariansim?" In *Contractarianism and Rational Choice*, edited by Peter Vallentyne: 15-30. Cambridge: Cambridge University Press.

Gintis, Herbert (2009). *The Bounds of Reasons: Game Theory nad the Unification of the Behavior Sciences*. Princeton: Princeton University Press.

Hamilton, W. D. (1964). "The Genetical Evolution of Social Behaviour I." *Journal of Theoretical Biology,* vol. 7: 1-16.

Hampton, Jean. (1986). *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.

Harsanyi, John. (1953). "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61: 434-435.

—— (1955). "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63: 309-321.

—— (1974). "Can the Maximin Principle Serve as the Basis for Morality?" *American Political Science Review* 69: 594-606.

Harsanyi, John and Reinhard Selten. (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Hobbes, Thomas. (1651 [1994]). *Leviathan*, Edwin Curley, ed. Indianapolis: Hackett Publishing.

Hubin, Donald. (1991). "Non-Tuism." *Canadian Journal of Philosophy* 21: 441-468.

Kalai, Ehud and Meir Smorodinsky. (1975). "Other Solutions to Nash's Bargaining Problem." *Econometrica* 43: 513-518.

Kliemt, Hartmut. (2009). *Philosophy and Economics I: Methods and Models*. Munich: Oldenbourg.

Luce, Duncan R. and Howard Raiffa. (1957). *Games and Decisions*. New York: Dover Publications.

Maschler, Michael, Eilon Solan, and Shmuel Zamir. (2013). *Game Theory*. Cambridge: Cambridge University Press.

Moehler, Michael. (2013). "Contractarian Ethics and Harsanyi's Two Justifications of Utilitarianism." *Politics, Philosophy, and Economics* 12: 24-47.

Nash, John. (1950a). "Equilibrium Points in n-Person Games." *Proceedings of the National Academy of Sciences of the United States of America* 36: 48-49.

— — (1950b). "The Bargaining Problem." *Econometrica* 18: 155-162.

— — (1951). "Non-Cooperative Games." *Annals of Mathematics* 54: 286-295.

Plamenatz, John (1973), *Democracy and Illusion*. London: Longman.

Plott, Charels R. (1996). "Rational Choice in Experimental Markets." *In Rational Choice: The Contrast between Economics and Psychology*, edited by Robin M. Hogarth and Melvin W. Reder. Chicago: University of Chicago Press: 117-43.

Rawls, John (1993), *Political Liberal*ism. New York; Columbia University Press,

— — (1971 [1999]). *A Theory of Justice, revised edition*. Cambridge: Harvard University Press.

— — (1974 [1999]). "Some Reasons for the Maximin Criterion." *John Rawls: Collected Papers*, edited by Samuel Freeman: 225-231. Cambridge: Harvard University Press.

— —(2001). *Justice as Fairness: A Restatement*. Cambridge: Harvard University Press.

Rosenberg, Alexander (1995). *Philosophy of Social Science*, 2nd edition. Boulder, CO: Westview.

Rubinstein, Ariel. (1982). "Perfect Equilibrium in a Bargaining Model." *Econometrica* vol. 50: 97–110.

Schelling, Thomas. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.

Shapiro, Ian and Donald Green. (1996). *Pathologies of Rational Choice: A Critique of Applications in Political Science*. New Haven: Yale University Press.

Schwartz, Norbert (2002). "Feelings as Information: Moods Influence Judgments and Processing Strategies." In *Heuristics and Biases: The Psychology of Intuitive Judgments*, edited by Thomas Gilovich, Dale Griffin and Daniel Kahneman. Cambridge: Cambridge University Press: 534-47

Sen, Amartya. (2009). *The Idea of Justice*. Cambridge: Harvard University Press.

Skyrms, Brian. (1996). *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

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

Smith, John Maynard. (1975). *The Theory of Evolution*, third edn. New York: Penguin.

Smith, Vernon (2002). "The Contrast Between Economics and Psychology." In his *Bargaining and Market Behavior: Essays in Experimental Economics*. Cambridge: Cambridge University Press: 7-25.

Sugden, Robert (1986). *The Economics of Rights, Cooperation and Welfare*. Oxford: Blackwell.

Thrasher, John. (2014). "Uniqueness and Symmetry in Bargaining Theories of Justice." *Philosophical Studies* 167: 683-699.

Thrasher, John and Gerald Gaus (forthcoming). "*Calculus of Consent.*" In *Oxford Handbook of Classics in Contemporary Political Theory*, edited by Jacob Levy. Oxford: Oxford University Press.

Trivers, Robert L (1971). "The Evolution of Reciprocal Altruism," *The Quarterly Review of Biology*, 66: 35-57.

Vanderschraaf, Peter. (2006). "War or Peace? A Dynamical Analysis of Anarchy." *Economics and Philosophy,* vol. 22 (2006): 243–79.

von Neumann, John. (1928). "Zur Theorie der Gesellschaftsspiele." *Mathematische Annalen* 100: 295-320.

von Neumann, John and Oscar Morgenstern. (1944). *The Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Weisberg, Michael (2007). "Who is a Modeler?" *British Journal for Philosophy of Science*, 58: 207-233.