The Use and Abuse of Formal Models in Political Philosophy

Robert Paul Wolff

April 25, 2011

Contents

List of Figures

List of Tables

Introduction

The purpose of these remarks is to introduce you to the technical foundations of a number of formal methods of analysis that have come to play a large role in the writings of philosophers, economists, political theorists, legal theorists, and others, and then to show you how these formal methods are misused by many of those theorists, with results that are conceptually confusing and quite often covertly ideologically tendentious. I am going to expound these materials carefully and with sufficient detail to allow you to master them and make your own judgments about the appropriateness of their use.

There are three distinct bodies of material with which we shall be dealing. Each has grown out of a different intellectual tradition, uses different methods of formal analysis, and finds different application by philosophers, political, theorists, and so forth. Quite often they are confused with one another, and my impression is that the people who use them frequently do not understand the distinctions among them, but we shall treat them separately.

The first body of material is Rational Choice Theory. When people talk about maximizing utility or calculating the expected value of an alternative or discounting an outcome by its risk, they are drawing on Rational Choice Theory.

The second body of material is Collective Choice Theory. When people talk about the paradox of majority rule or Arrow's Theorem or Pareto Optimality they are drawing on Collective Choice Theory.

The third body of material is Game Theory. When people talk about strategies or zero-sum or prisoner's dilemma, they are drawing on Game

Theory.

I am going to ask you to be patient, because this is going to take a while. By the time we are done, I may have written a short book. As we proceed, I will make some reading suggestions for those of you who wish to pursue the subject in greater depth, but everything you will need to know to follow my exposition will be contained in these pages.

The order of exposition is going to be as follows:

1. Some preliminary technical matters, principally concerning different kinds of orderings.

2. The elements of Rational Choice Theory.

3. The elements of Game Theory, maybe [if you have the stomach for it] including a formal proof of von Neumann's Fundamental Theorem concerning two person zero-sum games with mixed strategies. This will include some discussion on the misuse of Game Theory in nuclear deterrence theory.

4. A general discussion of misuses of the formal materials in treatments of so-called Free Rider problems and other matters. Depending on your endurance and interest, I may at this point discuss the use of formal models and materials in legal theorizing and other areas.

5. A formal analysis of John Rawls' claims concerning choice in the Original Position in *A Theory of Justice*.

6. The elements of Collective Choice Theory, including a formal proof of Kenneth Arrow's General Possibility Theorem, which is the central formal result in the field.[1]

[1]Final note in the original text: A REALLY, REALLY IMPORTANT REQUEST: I cannot see your eyes, so I cannot tell when they glaze over, either from boredom because I am going too slowly, or from confusion because I am going too fast. So I need to hear from you if either of those things is happening.

Part I

Technical Details

Chapter 1

Preliminary Technical Matters

1.1 Scales of measurement

Let us suppose that we have a finite set of discrete elements of any sort, which we will call $S = (a, b, c, \ldots, n)$.[1] The elements might be different amounts of money, different flavors of ice cream, different bowls of ice cream [not the same thing, of course], different candidates in an election, and so forth.

We may wish to impose an ordering on the set. The very simplest ordering we can impose is a nominal ordering, or a labeling. To each element, we assign a label or name [hence "nominal"]. Two or more elements may receive the same name, but each element receives only one name.

Such an ordering is said to be complete if every element in S is labeled. The ordering creates what are called equivalence classes, which is to say, subsets of elements all of which bear the same label or name. This labeling exhaustively and mutually exclusively divides S into subsets. Obviously, two elements are in the same equivalence class if and only if they have the same name. Every element is in one, and only one, equivalence class. With a nominal ordering, nothing more can be deduced from the labeling than the simple fact that two elements are in the same equivalence class if and only if they bear the same label. The essential fact about this very

[1] For more detail, see Stevens (1951), originally published in 1951 but re-issued and updated.

simple measure is that it is complete. Every element bears a label. For any two elements, either they are in the same equivalence class or they are not. Trivially, each element is in the same equivalence class with itself. Thus, every element is in some equivalence class.

The next step is to introduce a binary relation, R, over the set of elements. xRy is construed variously as meaning "x is equal to or greater than y," or "(someone is) indifferent between x and y or prefers x to y," or even "x is hotter than or is the same temperature as y," and so forth. All of these have the same formal structure.

Let us suppose the following two propositions are true for R and for S:

(i) for all x and y in S, xRy or yRx. This says that R is complete. Notice that from this, it follows that for all x, xRx. [Just as a trivial exercise, here is how we prove that xRx. Since for any x and y, xRy or yRx, take the case in which $x = y$. Then substituting, we have xRx or xRx, which is logically equivalent to xRx. That is the sort of baby logic steps we will be taking many of in what follows]. This property of an element bearing a relation to itself is called **reflexivity**, and a relation of which it holds is said to be reflexive.

(ii) for all x, y, and z in S, if xRy and yRz then xRz. This property is called **transitivity**, and it will turn out to be the single most important property of relations like R.

Just to be absolutely clear what we are talking about here, suppose we interpret the relation xRy to mean (someone) prefers x to y or is indifferent between x and y. Then (i) says that for any two members of the set S, the person in question either prefers x to y or is indifferent between them, or else prefers y to x or is indifferent between them. If this is still a bit puzzling, think of x and y as real numbers and R as meaning "is equal to or greater than." (ii) says that if the person in question prefers x to y or is indifferent between them, and also prefers y to z or is indifferent between them, then that person also prefers x to z or is indifferent between them.

A binary relation like R is said to establish a weak ordering on the set S. It is weak because it allows for indifference. Starting with the relation

R, we can also define a relation P on S, like this: xPy means xRy and not yRx. P here stands for "prefers," and a relation like P is said to establish a strong ordering on the set S. To get an intuitive handle on these very important little symbols, think of it this way. xRy says that x is at least as good as y, and maybe better. xPy says that x really is better than y [whatever "better" means here.] So R is weak and P is strong. Later on, when we come to Collective Choice Theory, we will be saying a lot about weak and strong orderings.

A relation, R, over a set, S, for which (i) and (ii) hold is said to be an ordinal ordering. In discussions of these matters in philosophy, economics, and political theory, it is often taken as a fundamental test of a person's rationality that his or her preferences exhibit at least an ordinal ordering over all available alternatives.

Some economists, using what is called a theory of "revealed preference," even argue that everyone must have a preference structure that at least satisfies the first condition, and thus is complete, because confronted with any two alternatives, x and y, a person will either choose one, thus showing that she prefers it to the other, or else will be indifferent between the two. But that, I will argue much later, is a covertly tendentious thesis made more plausible by the formalism. Think *Sophie's Choice.* [I.e., first you force a woman to choose which of her two children you are going to kill, and then you say, "So, that shows she prefers the one to the other." I am going to have a good deal to say about this sort of thing down the line.]

By the way, "ordinal" because the ordering merely establishes which of the elements is first, second, third, fourth, etc. according to the relation R, and these are what are called "ordinal numbers."

It may not be obvious at first glance, but preference structures do not always exhibit transitivity, and hence are not even ordinal. Indeed, the casual assumption of transitivity is actually an enormously powerful and simplifying assumption.

Let me give an elementary and non-controversial example here, and save the controversial examples for later. All of us, I assume, have had our eyesight checked at the optometrist's office. You shut one eye, the room is

darkened, and you look through a complicated gadget at a chart of rows of letters, each line smaller than the one above. The doctor flips lenses in front of your open eye, and asks: "Which is clearer, one, or two?" Sometimes you can see a difference, and sometimes you just say, "They are the same." The two lenses may actually have different degrees of magnification, but the difference is simply too small for you to notice. Experimental psychologists say that the difference between the two is then below your "minimal discriminable difference," or MDD. Now, it is obvious that with a little work, the optometrist could line up a series of lenses, each successive pair of which falls below your MDD, but the first and last of which are clearly discriminable. If we interpret R in this case to mean "is clearer than or is equally as clear as," it would be true that for any adjacent pair, m and n, mRn and nRm, but for the first, a, and the last, q, it would not be the case that aRq and qRa. In other words, the relation "is clearer than or equally as clear as" would not be transitive.

The same thing might manifestly be true of someone's preferences. What all this means is that it is very powerful and quite probably false to assume that someone has a transitive preference ordering over a set of available alternatives. But people who use this sort of formalism almost never realize that fact. Indeed, it is quite often the case that people introduce this formalism without even feeling any need to say that they are assuming transitivity. This is a simple example of what I mean when I say that the formalism can conceal powerful and dubious assumptions.

Ordinal preference orders encode the order of someone's preferences, but not the intensity of that preference. Compare voter A with voter B in the 1992 presidential election. Voter A is a fanatic George H. W. Bush supporter. She doesn't really like either Clinton or Perot, but despite Perot's kookiness, prefers him by a hair to Clinton. Voter B is torn between Bush and Perot, neither of whom he loves, but he finally decides to go with Bush. He hates Clinton and wouldn't vote for him even if Mao Tse-Tung were the alternative. These voters have identical ordinal preference structures: Bush first, Perot second, Clinton third. That is all you need to know to figure out how they will vote, but obviously for all sorts of other purposes this ordinal preference ordering fails to embody a great deal of

important information. In particular, this ordering will not tell you how either voter might behave in other political contexts besides voting, such as donating money, working for a campaign, lobbying, and so forth.

This is as good a place as any to call into question the easy assumption that the possession of a complete ordinal preference structure is the most elementary test of one's rationality. A great deal is at stake here, much more than you might think. Let us start slow. The theory of rational choice has its roots in analyses of gambling behavior, of economic behavior, and—to some degree—of political behavior. Now, when we are talking about the way in which professional gamblers decide how to play their cards or place their bets, it makes sense to assume that they can define a complete preference order over the available alternatives. That is to say, the various possible outcomes offered by a gambling game are plausibly described as commensurable with one another. The outcomes are, after all, simply different wins or losses of amounts of money. The same is true of people engaged in economic activities. But these are relatively limited and specialized arenas of human activity. There are many other arenas in which it is not so obvious that rational individuals have complete preference orders over available alternatives.

Consider, as an example, the terrible choice presented to the central character in William Styron's novel *Sophie's Choice*. [I know the story from the movie of the same name, starring Meryl Streep.] A Gestapo *gauleiter* overseeing the loading of Jews onto trains taking them to the death camps offers Sophie a choice. She may save one of her two children from certain death, but she must choose which one will survive. His posing of this choice is clearly an act of satanic sadism. There are two ways of thinking about this situation. The natural, and I suggest, rational way to think about it is as a tragedy in which a woman is presented with a terrible situation that will destroy her life no matter what she does. To choose either child is impossible. To fail to choose one is to condemn them both to death. Religion may have something useful to say about this situation. Literature may. Perhaps nothing can. But for sure Rational Choice Theory is no help. But Rational Choice Theory says that she must have some preference order or other over the three outcomes, and her choice, whatever it is, reveals

that preference.

Let me put this in a summary fashion, and ask you to think about what I say. Perhaps later on we can discuss it. The assumption of a complete ordinal preference order is presented in the literature as an innocuous premise that gets the more complex and interesting arguments going. But in fact it is, covertly, a highly questionable proposal to extend a form of economic rationality into areas in which it arguably does not belong. By accepting the formalism, someone unwittingly buys into this powerful encroachment of the economic into arenas of human experience in which it has no place. Imagine coercing a man into acting dishonorably, and then saying that his agreement reveals exactly what price he places on his honor. It would be more true to the human reality to say that by this act of coercion, you have besmirched his honor, which henceforth is worth nothing to him. The outcome of the choice you forced on him is not a rational choice but shame.

The defining characteristic of capitalism is the reduction of all human activity to market relations. Too often, Rational Choice Theory functions as a covert and seductive rationalization of the capitalist *ethos*, which then seems, because of the apparent neutrality of the formalism, to be equivalent to rationality *tout court*.

It is not necessary to limit ourselves to complete orderings—orderings which establish the individual's or society's preference for any two alternatives whatever. We can also define partial orderings, and these have in fact played an important role in Economics and other disciplines. I will only say a few words here, and return to this subject down the line. The *Sophie's Choice* example has shown us that sometimes individuals cannot say, for two alternatives, which one they prefer. It is not that they are indifferent between the two. The two are simply, in their minds and hearts, not comparable. How many lives is it worth to save the only score of Bach's B Minor Mass? The question makes no sense to us, no matter what phony scenarios we cook up in a philosophy essay.

A similar problem arises when we are trying to compare different social distributions of wealth. If Situation B offers everyone more wealth than Situation A, then we can be pretty sure there will be unanimous

agreement that B is better than A. Indeed, if people are willing not to be envious of what others get, then we might be able to secure unanimity for the proposition that B is better than A if B offers everyone at least as much as A does, and offers at least one person more. [Why begrudge her the extra if it isn't coming out of your share?] But what about the case in which B makes some people better off and others less well off than they were in A? There may just be no answer in that case.

Thanks to Vildredo Pareto [1848 - 1923], when B makes everyone better off than they were in A, we say that B is Pareto Preferred to A. Obviously, if B is Pareto Preferred to A and C is Pareto Preferred to B, then we should expect that C will be Pareto Preferred to A. So this Pareto or Unanimity ordering is transitive but not complete. If some way of distributing things is such that there is no alternative distribution that is Pareto Preferred to it, then we say that it is Pareto Optimal. Don't be misled by the enticing sound of the word "optimal." If we assume that everyone has positive marginal utility for money, so that taking even a little bit away from someone makes her less well off, then a social distribution that gives everything to one man and nothing at all to anyone else is Pareto Optimal, because any re-distribution will involve making at least one person worse off, namely the person who had everything and now has slightly less. In case you are wondering why this matters, I will just point out that when economists describe a market as efficient, they mean that it produces a Pareto Optimal outcome. Not too heart warming.

So much for ordinal preference orders, at least for the moment. Now things get somewhat more complicated, but also a good deal more important. The next step up, after nominal and ordinal orderings, is cardinal orderings. Since this is going to require a little technical work, let me first explain what is at stake. Both Rational Choice Theory and Game Theory [but not Collective Choice Theory] involve talking about people doing something called "maximizing expected utility," or "discounting the value of an outcome by its risk" and so forth. These calculations require that we be able to assign cardinal numbers, or magnitudes, to different outcomes or alternatives, and that we be able then to do things like adding them, subtracting them, multiplying them by other numbers, etc. Now, you can-

not add or subtract or multiply or divide ordinal numbers. It makes no sense to ask, "Is Second the average of First and Third?" in the way that you might ask "Is 2 the average of 1 and 3?" If you have ever been involved in trying to work out a system to decide which team in a track meet has won over all, or which country has won over all in the Olympics, you will understand this. Does a whole raft of silver and bronze medals count for more or for less than a small pile of gold medals? Are a gold and a bronze equal to two silvers? The questions are meaningless. To carry out any of these calculations, you need an interval scale, also called a cardinal ordering.

An interval scale is an assignment of numbers to the elements of an ordinal ordering in such a way that the intervals are equal [hence "interval scale." This is actually a gross simplification of the correct definition, but I don't want to scare people away, and this will suffice.] A good example is the Fahrenheit temperature scale. The elements here are, let us suppose, readings provided by a thermometer of the temperature of different bodies of water. We can first impose a nominal ordering by grouping together the bodies of water that are [or maybe feel] the same temperature. We then impose an ordinal ordering by arranging the equivalence groups in a hierarchy from hottest to coolest. Thus far, all we have is the information that this body of water is hotter than that one, or maybe that this body of water is at least as hot as that one [i.e., weak rather than strong ordering]. Now, suppose we can actually answer the following question for any four bodies of water, a, b, c, and d: Is the difference between the temperature of a and the temperature of b at least as large as the difference between the temperature of c and the temperature of d? Notice I said any four bodies of water. In other words, I am asking about intervals of temperature, not just temperatures. If I have enough information to answer that question for any tetrad of bodies of water, then I can define a cardinal measure of temperature. I can say, for example, using the Fahrenheit scale, that the difference or interval between fifty degrees and sixty degrees is the same as the difference or interval between twenty degrees and thirty degrees. So it makes sense to say, "It is ten degrees cooler today," regardless of what the temperature was yesterday.

We are here performing an arithmetic operation on the labels assigned to the elements of the set [namely subtraction]. But you cannot perform arithmetic operations on ordinal numbers. For that you need cardinal numbers [i.e., real numbers] like 1, 2, 3, and 4. So, this sort of scale is called a cardinal scale.

This right here is one of the most important things I am going to explain, so if it is not clear, I expect to hear from you.

The last step, which is only important for a few purposes, is to define what is called a ratio ordering or a ratio scale. A ratio scale is just like a cardinal scale but with one thing added: with a ratio scale, we have enough information to say, for any four elements of our set, a, b, c, and d, whether the ratio of a to b is equal to or greater than the ratio of c to d. Or, going back to the symbolism we used above, whether $\frac{a}{b} R \frac{c}{d}$. Now a little experimentation will show you that a ratio scale requires that you be able to identify some point as the zero point, or origin. A Fahrenheit temperature scale is not a ratio scale. The zero point in the Fahrenheit scale is chosen arbitrarily. Therefore, it makes no sense at all to say that in a Fahrenheit scale, the ratio of twenty degrees to ten degrees is the same as the ratio of eighty degrees to forty degrees. [For those of you who know some Physics, that sort of statement does make sense in a Kelvin scale of temperature, where the zero point is what is called absolute zero.]

Why am I going on about this? Well, for one reason, because it will turn out, way down the line, that without knowing this stuff you cannot understand what a zero-sum game is.

1.2 Transformations

Now we are going to talk about transformations. Technically, a transformation is a one-one mapping of a set onto another set, but we can just think about a transformation as a rule for assigning new labels or numbers to a set of elements.

1.2.1 Permutation

A permutation is a re-assignment of the labels attached to the elements that preserves their grouping into equivalence classes. Initially, you will recall, we attached labels to the elements of our set, S. Two elements that got the same label were then in the same equivalence class. So if we label people by their last names, all the Millers go together, all the Tailors go together, and so forth. We could now relabel everyone, say by translating their names into another language [so that all the Millers become Muellers, and all the Tailors become Schneiders.] That would change everyone's name, but it would not change the groupings into equivalence classes. All the Millers were together before, and they are still together now that they are all Muellers. What is more, no two people who were in different groups before are in the same group now. The official jargon for this state of affairs is that the labeling is invariant under a permutation.

1.2.2 Monotone transformation

A monotone transformation is a re-labeling that preserves an ordinal ordering. Suppose we take the items labeled first, second, third, and fourth, and now label them fifth, eleventh, nineteenth, and fortieth. No information in the original ordering has been lost, and none has been gained. In the formalism of the relation R, for any two elements a and b in S that have been relabeled a' and b' respectively, aRb if and only if $a'Rb'$. So the ranking has not been changed by the transformation. Again, the official way to say this is that the ordinal ordering, R, is invariant under a monotone transformation.

1.2.3 Linear transformation

A linear transformation is a transformation that preserves a cardinal ordering. A linear transformation of a relation R on a set of elements $S = (a, b, c, ...n)$ is a relabeling of each element a in S such that the new label, a' equals the old label a times some constant plus another constant. Or: $a' = aq + r$. This is called a linear transformation because the

expression ($a' = aq + r$) is the formula for a straight line drawn on the a and a' axes. A little elementary algebra will show that this transformation preserves an interval scale or cardinal ordering on the elements of S. Remember: This means that it preserves equality of intervals between pairs of elements. Here is how we prove this:

Take four elements, a, b, c, and d, such that $(a - b) = (c - d)$. Now impose a linear transformation on S. That means:

$a' = aq + r$

$b' = bq + r$

$c' = cq + r$

$d' = dq + r$.

Notice that we have imposed the same linear transformation on each element. In other words, the constants q and r are the same in each case.

By hypothesis $(a - b) = (c - d)$

Substituting the transformed labels, we get $(aq + r - bq - r) = (cq + r - dq - r)$

or: $(aq - bq) = (cq - dq)$

Dividing both sides by q, we get $(a - b) = (c - d)$ Ta Da!

Figure 1.1 has a graph with two lines on it, showing you that the line $a' = qa + r$ ($A' = 0.5a + 2$, represented by a solid line) cuts the a' (i.e., vertical) axis at r (2) and cuts the a (i.e., horizontal) axis at $-r/q$ (-4). The dotted line, on the other hand, represents $a' = -qa + r$ ($a' = -0.5a + 6$), which cuts the vertical axis at r (6) and the horizontal one at $-r/q$

$(-(-6/0.5) = +12)$. So long as you re-label the elements of S so that the labels satisfy the equation $a' = qa + r$, for any a in S, it makes no difference which set of labels you use, because they all encode the same information. This is what it means to say that a cardinal ordering is invariant under a linear transformation.

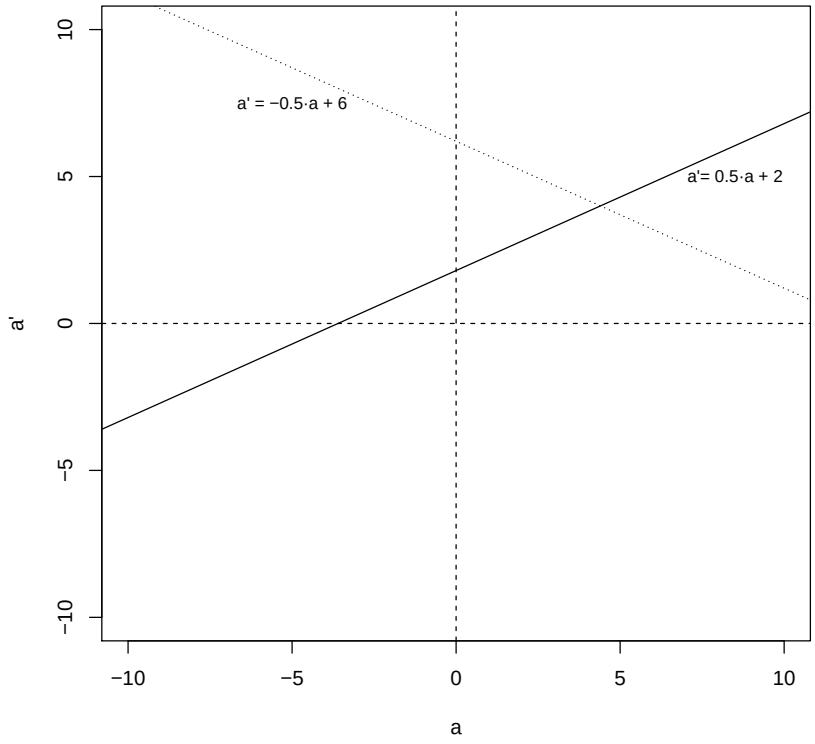


Figure 1.1: An example of representation of $a' = qa + r$ and $a' = -qa + r$.

A linear transformation does two things. It changes the size of the intervals [but not the equality of different pairs of intervals], and it changes the zero point. The classic example of a linear transformation is the formula for converting temperature from Fahrenheit to Centigrade. The formula, as everyone knows who travels to Europe and wonders whether to wear a sweater or not, is F degrees = 9/5 Centigrade degrees + 32. So, if *Le Monde* says it is going to be 20 degrees today in Paris, that means 9/5(20) + 32 or 68, so no sweater. Zero degrees in centigrade is the temperature at

which water freezes, but in Fahrenheit, that is 32 degrees. And so forth. Each degree Centigrade is equal to 9/5 Fahrenheit degrees. Why does this matter? Once again, it will turn out to be crucial when we come to give a correct definition of a zero sum game, and for many other purposes besides.

1.2.4 Ratio transformation

Finally, and uninterestingly, a ratio transformation is a transformation of the form $a' = qa$. This the formula for a line that goes through the origin of the graph. The transformation just changes the size of the intervals but does not change the zero point. And obviously, $a'/b' = qa/qb = a/b$. This one doesn't matter much, but I put it in for completeness' sake.

O.K. We have nominal, ordinal, interval, and ratio scales, and we have permutations, monotone transformations, linear transformations, and ratio transformations.

Now, as Portnoy's analyst says in the last line of the novel, let us begin.

Pedagogical Note: This exposition is intended for people unfamiliar with the material. I am trying to explain things slowly and clearly, without inside baseball allusions to sophisticated interpretations. I want each step to be completely clear to anyone who is following along. Maybe at the end I can discuss some of the trickier mathematics. I do not believe that makes any difference to the applications of this material in law, political science, etc.

Part II

Theoretical Details

Chapter 2

The elements of Rational Choice Theory

Gamblers have always known that in deciding how to place your bets, it is essential to take into account both how much you can win or lose and how likely you are to win or lose. One bet may offer a chance for an enormous payoff [a national lottery, say] but very little chance of winning, while another bet offers a pretty good chance of a small gain. "Enormous" and "small" are not much help, and neither are "little" and "pretty good." How should you evaluate the relative attraction of two bets? How should you decide what is a reasonable entry fee for playing a gambling game?

For a very long time, one standard answer has been to calculate what is called the mathematical expectation of a gamble. This is the size of the possible gain discounted, or multiplied by, the probability of winning. If a gamble pays winners $100 and players have a 10% chance of winning, then the mathematical expectation of the gamble is (100)(.1) = $10. A prudent gambler will pay no more than ten dollars to play that game.

What exactly does it mean to say that the mathematical expectation of the game is $10? Well, one thing it does not mean is that there is any chance at all of winning exactly ten dollars. If you play this game, there are only two possible outcomes: either you will win one hundred dollars or you will just lose your entry fee. The standard answer is that if you play the game over and over again, your average winnings will tend to cluster

more and more tightly around ten dollars. This is the sort of calculation that casinos and casino gamblers make. The games are all set up so that over the long run, the House tends to average a small gain per play. That is why casinos make money.

There are two problems with this explanation, both of which play an extremely important role in the application of this formalism to political theory, legal theory, military strategy, and so forth. The first problem is that not all of life is a casino, and in some situations we are not presented with the realistic possibility of "playing" over and over again. Think nuclear war. The second problem is that very long runs of losses are possible, even in a game with an attractive mathematical expectation, and there is always the chance, as you play again and again, that you will run out of money before things "average out."

Are there any other rules someone might propose for making "rational" choices? Indeed there are, but before we talk about any of them, we need to do a good deal more technical preparation.

Notice that to carry out a calculation of mathematical expectation, there are two things you must know: the precise value of each outcome, be it a win or a loss, and the precise probability of each possible outcome. If you do know both of these things, then you are said to be making a decision under risk. If you do not know the probabilities of the several outcomes, but you do know the precise value of each possible outcome, then you are said to be making a decision under uncertainty. Game Theory analyses decision under uncertainty. The Original Position in Rawls' theory is a situation of decision under uncertainty. Maximization of expected utility presupposes decision under risk.

If you are not sure what all the possible outcomes of a choice are, or you do know what they are but do not know what value you place upon them, then these theories have nothing to tell you about how you should make decisions, even though that is a pretty good description of the fix we usually find ourselves in the real world. As my favorite Rational Choice Theorist, Donald Rumsfeld, liked to say, during his glory days when the Iraq invasion was going the way he wanted it to, there are the knowns and the known unknowns, and then there are the unknown unknowns.

At this point, enter Nicolaus Bernoulli [1687-1759], cousin of the Bernoulli whose work helps to explain why airplanes fly. It seems that in St. Petersburg, which was, in the 18th century, a popular watering hole for European aristocrats, a game was being played by skilled and knowledgeable gamblers whose betting behavior seemed to contradict the well-known rule of maximizing mathematical expectation.

The gamble was this: a fair coin is tossed again and again until it comes up heads. You receive, as your winnings, a number of ducats equal to 2 raised to the $(n-1)$ power, where n is the number of the toss on which heads appears. [Ducats because you couldn't spend rubles anywhere but in Russia, and there are only so many sets of nested Russian dolls you can buy.] So if the coin comes up heads on the first toss, you get 1 ducat. [Why? Because in this case, $n = 1$, and $1 - 1 = 0$, and 2 to the zero power is 1] if the coin does not come up heads until the second toss, then you get 2 ducats. [Why? Because in this case $n = 2$, and $2 - 1 = 1$, and 2 to the first power is 2.] If heads comes up on the third toss, you get 4 ducats, on the fourth toss 8 ducats, and so on ad infinitum. The question for the gamblers was: How much should I be willing to pay to enter the game? Now the classic answer, well understood was: Pay any amount up to the mathematical expectation of the game. So all that was necessary was to calculate the mathematical expectation of the game.

You calculate the mathematical expectation of the game by taking the payoff of each possible outcome [heads on the first toss, heads not until the second toss, heads not until the third toss, etc.], discounting that payoff by the probability of that outcome, and then adding up all *th* discounted payoffs. That is the mathematical expectation of the gamble.

O.K. There is a 1/2 chance that heads will turn up on the first toss, and $(n-1)$ in this case is 0. 2 to the 0 power is 1. So, discounting the value of the payoff, one ducat, by the odds, one-half, the value to a gambler of heads turning up on the first toss is 1/2 ducat. There is a 1/4 chance heads won't turn up until the second toss. If that happens, then $(n-1)$ is $(2-1)$ or 1. So the payoff is 2 ducats, and the odds of getting it are 1/4, so the expected value is again 1/2 ducat. A little thought or a little experimentation will you show that every payoff of this endless series of

terms is 1/2 ducat. And the sum of an infinite number of 1/2s is infinite. In other words, the MATHEMATICAL EXPECTATION of the gamble is INFINITE!

So according to the rule that everybody accepted, namely evaluating a gamble as equal to its mathematical expectation, a gambler should be willing to pay any amount of money he can lay his hands on to enter the game. Or, as Bernoulli rather quaintly puts it, a gambler should refuse to accept any amount of money that another gambler offers him for his ticket to play the game. But that is crazy! As Bernoulli observed in St Petersburg, experienced gamblers would not dream of doing anything like that. How to explain this apparently irrational behavior?

Bernoulli's answer was that the gamblers, contrary to popular opinion, were not trying to maximize their money winnings. Instead, they were trying to maximize the utility they got from their money winnings. I am going to continue with Bernoulli's analysis, and put off for a few paragraphs a deeper look at his solution to the puzzle, but let us be clear right here what he is saying. According to Bernoulli, the gamblers know what the mathematical expectation of the gamble is, in money, and they also know how much "utility," whatever that is, those amounts of money will give them. We will come back to this very soon. Let us continue. Drawing on the theory of logarithms, which had been around for about a century, thanks to John Napier, Bernoulli decided that there was a logarithmic relationship between the amount of money a gambler won or lost and the amount of utility he got from that money. Indeed, Bernoulli claimed to know the formula. It was, he said:

$$u = b * log(a + D)/a$$

Where a is the amount of one's fortune before the gamble, D is one's winnings from a toss of the coin, u is the utility the gambler gains from the winnings, D, and b is a constant to be determined empirically—by observation, one supposes. Bernoulli then demonstrated mathematically that if a pauper was given an entry ticket to the game [a pauper is someone with no previous fortune, so $a = 0$], the rational thing for him to do would

be to sell that ticket for as little as 2 ducats. For an ordinary gambler, who presumably already has some money in his pockets, the ticket would not be worth even that much.

There is a whole lot to say about this little story, so settle down. First of all, where on earth did Bernoulli get that formula from? The answer is pretty clear. He invented it, because he knew how to manipulate it mathematically to get the answer he wanted. For our story, it doesn't really matter, because that formula is going to disappear pretty soon, and never reappear in this blog. I just put it in because once in your life you should know what Bernoulli said. Pretty clearly, he decided that was the shape of the utility function because he knew how to solve that formula for $D = 0$. Here are the really important things that can be said:

1. If you plot that formula on a graph whose x axis is the amount of money won and whose y axis is the amount of utility gained from that money, you will find that the line rises sharply at first and then bends more and more toward the horizontal, so that as it goes out infinitely to the right [representing longer and longer streaks of all tails before a head shows up], it approaches the horizontal asymptotically. So the total value of the entire gamble in utility approaches some finite amount. Each additional half ducat one gains by yet another tails yields less additional utility than the half ducat gained by the previous appearance of tails. This is called having declining marginal utility for money [or at least for ducats], and if you can grasp this idea, you can understand most of modern Economics. For all manner of purposes, economists routinely assume declining marginal utility for money. The idea is that although the next dollar is probably worth a good deal to you if you are of modest income, if you keep getting dollars, after a while each additional one will be worth less and less to you. If people's utility functions are nice and regular like this—continuous, monotonically increasing, with declining marginality—then we can use the calculus to do all sorts of nifty things that make economists feel really good about themselves. But the fact is that people's utility functions, assuming they have them,

are much less conveniently smooth than that. They are lumpy, they go backwards, and in all sorts of ways are not amenable to the simple manipulations that the calculus allows us when dealing with what are called continuous functions. Leonard Savage and Milton Friedman, a long time ago, published a famous little paper in which they pointed out that since the same people often buy both lottery tickets and life insurance [in the first case paying for the privilege of risk and in the second case paying to avoid risk], their utility functions cannot exhibit monotonic declining marginal utility.

2. The second point is equally important, but not usually mentioned. So long as you are talking about the money payoffs, you are in the realm of the public, the objective, the easily measurable. But once you shift to talking about utility, all of this easy publicity disappears. Each person has his or her own utility function relating money to utility, and there is no reason at all to suppose that any two people have identical utility functions. Furthermore, for all manner of well-known philosophical reasons, it is impossible to compare one person's utility with another person's utility [basically because you cannot see into another person's mind]. And since a person's utility function, even if it is a cardinal function, is invariant under a linear transformation, there is no way of knowing whether one person's units of utility are bigger or smaller than another person's units, nor is there any way of knowing the relation between the zero point of one person's scale of utility and the zero point of another person's scale. Both of those [scale and size of unity] are arbitrary. It is exactly as though you found two thermometers using different scales of temperature, and there was no way of sticking them into the same bucket of water to discover the conversion formula. [Now you begin to see why I went through that technical stuff earlier]. Everyone has had the experience of thinking that some guy is a sissy because he cannot stand a little pain, even though his wife went through childbirth without an epidural. But suppose he replies that he is more sensitive to pain, and suffers more from something that others find

bearable. Not even modern neurophysiology can determine whether that is true or false. Recall the aristocrats who claimed that peasants have courser sensibilities and therefore suffer less from sleeping on rocks than a princess does from having a pea under her mattress. The appropriate response to that claim is not Rational Choice Theory. It is the guillotine. For all of these reasons, it never makes sense to try to add together the utilities of two different people, unless some very special conditions are present [see discussion below of the concept of a zero sum game.]

3. The third point is that Bernoulli gives us no way to figure out what someone's utility function is, and neither do most of the people who followed after him and adopted the practice of talking about utility functions. When we get to Game Theory, I will go through the rather complex set of premises that Howard Raiffa and Duncan Luce lay down in their invaluable book, *Games and Decisions*, from which one can deduce that someone has a cardinal utility function invariant under linear transformations. You will see that it is a huge leap of faith to suppose that people have cardinal utility functions. It is even a leap of faith to suppose that they have utility functions at all.

4. But the big problem is, Bernoulli does not tell us what utility is, and neither do any of the people who follow him. This is a huge subject, and I can only scratch the surface of it. Here goes. The word "utility" means "usefulness," which immediately raises the question, useful for what? Intuitively, we do not think of pleasure useful. It is, as we say, an end, not a means. David Hume, in his great work *A Treatise of Human Nature*, speaks of things that are "useful or agreeable to ourselves or others," clearly implying that a thing that is useful is useful for getting something else that is agreeable. In the modern discussions of what is called "utility theory," or "the theory of expected utility," this distinction is simply ignored, and utility is treated as somehow equivalent to pleasure. This confusion or unclarity was made worse by the ethical and political theorists—James Mill, John Stuart Mill, Jeremy Bentham—who asserted that an

act is right only insofar as it produces the greatest happiness for the greatest number, and then called that view Utilitarianism.

5. If you put these various comments together, here is the sort of problem you arrive at: Rational Choice Theorists take it for granted that the rational thing to do in any situation is to choose the alternative that maximizes expected utility. This presupposes four things, not one of which they can plausibly argue for: First, that we know all the possible outcomes in a situation and their probabilities; Second, that each of us has, and has access to, a cardinal utility function invariant under linear transformations that takes as its argument an outcome and has as its value the measurable quantum of utility that outcome will yield; Third, that we know what quality, experience, or state of mind we are referring to when we speak of a quantum of utility; and finally, that we should choose in accordance with the principle of the maximization of expected utility even in situations in which there is no realistic opportunity to repeat the choice endlessly many times so as to generate a series of outcomes. The elegance of the mathematics seduces Rational Choice theorists and others into sliding past all these serious issues so that they can get to the fun stuff of playing with the mathematics.

So much for the easy stuff. Now let's say a word or two about more complex issues that play a very important role in criticizing the application of Rational Choice Theory and Game Theory to military strategy and nuclear deterrence. We have been talking about maximizing expected utility, as though it were obvious that two alternative actions or strategies or choices with the same expected utility are equally worthy of being chosen. But a moment's thought shows that this assumption is, at the very least, questionable.

A simple example will make the point. Suppose I am presented with the opportunity to play either of two games. The first offers a coin toss, with heads winning me an amount of money for which I have very great utility, and tails losing me an amount of money for which I have exactly the same utility. [I have to define the game in this clumsy way, remember,

because it is, by hypothesis, utility and not money that I seek to maximize. Given the shape of my utility function, it might be that for me, a million dollars gained is equal in utility to the one hundred thousand dollars I already have. So this might be a coin toss game that wins a million if heads comes up and loses a hundred thousand if tails comes up.] The expected utility of this game is, by construction, zero. [1/2 times the utility to me of a million dollars 1/2 times the utility to me of a hundred thousand dollars, where, by hypothesis, those two amounts of utility are equal.] The second game consists of the game master simply saying to me, "You neither win nor lose anything." The expected value of this game is also zero.

Now, the theory of rational choice says I should be indifferent between these two games. There is, according to my calculations of expected value, no reason to prefer one to the other. But in fact, as I think is obvious, some people would clearly prefer to play the first game, while others [myself included] would prefer to play the second. This is not, let me emphasize, because I value the million I might win less highly than the one hundred thousand I already have [assuming that I have it, hem hem]. If that is true, then just adjust the amounts until the utilities are equal, wherever in dollar amounts that balance lies. [There must be such a pair of amounts, by the way. That is one of the implications of the assumption that my utility function is reflexive, complete, and transitive.]

Intuitively [and correctly], the explanation for the varying ways in which different people would rank these two games is that people have different tastes for risk itself, independent of their calculation of expected value. Some people like to take risks, and others are risk averse. Take me, for example. I don't like risks. Suppose I decide [who knows how?] that fifty dollars is worth twice as much to me as twenty dollars [because I have declining marginal utility for money]. If you offer me a sure twenty dollars or a fifty percent chance of getting fifty dollars, I am as likely as not to take the sure twenty, because I just don't like risk. I know that the mathematical expectation of the risky alternative is (1/2 x 50) or 25 dollars. And since I have positive, albeit declining, marginal utility for money, I prefer $25 to $20. Even so, I will take the sure $20. I have better things to do with my life and I just don't like risk.

This problem was the subject of a fascinating debate fifty years ago or so between the French economist Maurice Allais and the *emigré* Ukrainian economist Jacob Marschak. Allais argued the point I have just been making. Marschak replied that the problem of attitudes toward risk itself could be got round by changing the nature of the set, S, of alternatives over which a subject is asked to express preferences. Instead of a set of outcomes, or payoffs as they are frequently referred to in the literature, you can present the subject with a set of what Marschak called prospects, which are total future states of affairs. Since a prospect includes the pattern of risk involved in the making of a choice, preference for risk itself can be built into the utility function, thus getting around the fact that people have different tastes for risk independently of their attitudes toward the various outcomes that may result from a gamble.

This response is correct, and can easily enough be handled mathematically, but it misses a deeper point that is, I believe, fundamental. The whole purpose of introducing the concept of a utility function and the associated process of maximizing expected utility is supposed to be to provide a chooser with a definite and calculable method for making a decision confronted with alternatives, based only on the chooser's utility function. In effect, the theory says to someone making a choice, "If you know how you feel about the outcomes [your utility function] and if you know the probabilities [the premise of choice under risk], then this method will allow you to calculate what it is rational for you to do, even when it is unclear to you what that is." If this claim can be sustained, then the method of expected utility maximization is a very powerful aid to rational choice. But if it is necessary to shift to a utility function defined over total prospects, then all of the power and usefulness of the rule of expected utility maximization is lost. This may not be clear when one is awash in formalism and symbolism, but if you remind yourself what those symbols actually mean, and do not let yourself be beguiled by the spiffiness of the mathematics, then the force of Allais' objection is clear [in my opinion].

There are also a number of more subtle points relating to the construction of the utility function. In order for a cardinal utility function to be constructed from someone's preferences, it is necessary that all of

the outcomes in the set S be commensurable with one another. That is, it must be possible to represent the subject's preferences by an assignment of finite cardinal numbers, so that for any three alternatives a, b, and c in S, there is some probability p for a such that:

$$b = pa + (1 - p)c.$$

Since $p + (1 - p) = 1$, the expression on the right of the equation says that it is certain that either a or c will happen. The equation says that there is some way of adjusting the probabilities so that the subject is indifferent between outcome b and the gamble of a or c with the probabilities p and $(1 - p)$.

But it may be that one of the possible outcomes is, in the eyes of the subject, so much worse than any of the others [for example the subject's death] that there is no probability of that outcome, however small, that the subject is willing to risk. Alternatively, there might be one outcome so much better, in the subject's view, that there is nothing else you can offer the subject to compensate her for losing even the tiniest bit of her chance of getting it [for example, eternal salvation]. If either of these is the case, then the subject does not have a cardinal preference ordering, but instead has what is called a lexicographic preference ordering. Since this will come up later, a word about lexicographic preference orderings.

When we alphabetize a group of words [hence "lexicographic"], we put first all the words that begin with the letter *a*, regardless of what the subsequent letters in the word are. We put *azure* before *bad*, because kit starts with the letter *a*, even though the *z*, the *u*, and the *r* in *azure* come relatively late in the alphabet, whereas the letters *a* and *d* in *bad* come early. The earliness of *a* and *d* does not, as it were, compensate for the fact that *b* comes after *a*, nor does the lateness of *z*, *u*, and *r* count against *azure* in the alphabetizing. In other words, we are not assigning numbers to the letters and then arranging the words in the order of the sum of the letters in them [as medieval Hebrew scholars did in the *Kabbalah*]. Arranging a set of alternatives in this fashion, with one or more alternatives being, as we say, "lexicographically prior to" the others, yields a lexicographic

ordering of the set.

Keep this in mind, along with everything else I am telling you. It will turn out to play a role in my criticism of the application of Game Theory to military strategy by deterrence theorists, and also will turn out to pose problems for Rawls.

Well, that was fun. Now let us discuss an even hairier problem that actually played a very important role in decisions made by the Defense Department in the 1960s about the construction of the command and control systems for America's nuclear weapons [we are talking serious stuff here, folks.]

As I explained in my blog, the enormous destructive power and revolutionary character of nuclear weapons forced America's military planners to turn for advice to economists, psychologists, mathematicians, and philosophers. Very quickly, a number of these think tank defense intellectuals began to worry about the following problem. If the Soviet Union should be so foolhardy as to launch a first strike nuclear attack on America, it might, as part of this attack, target Washington D.C. In an instant [quite literally, in an instant] every decision maker of any constitutional authority in Washington might go up in a mushroom cloud. At the same time, almost certainly communications among those remaining alive would be disrupted by the effects of the explosions occurring across the country. The nuclear submarines carrying missiles with multiple separately programmable warheads would still be functional, presumably, but they might be out of contact with whatever remained of the military or civilian high command.

It was clear to the defense intellectuals that two things needed to be planned for and implemented. First, a physical system of backup communications and control of warhead delivery systems had to be put in place now, so that even after the incineration of the president and his so-called black box, it would be physically possible to use the remaining missiles, if that what was what it was decided to do. Second, a set of standing orders had to be promulgated now, directing officers [or even enlisted soldiers] still in possession of usable nuclear weapons to carry out whatever orders it was decided, *ex ante*, to give them. Because of the instantaneity

and scope of nuclear destruction, it was clear that those responsible for making decisions about the use of nuclear weapons could not wait until after the attack to deliberate and decide. The relevant people might not survive the attack, and even if they did, they might not be in a position to issue orders that could be received. The response had to be planned for in advance, if there was to be a response at all.

To the defense intellectuals, who were accustomed to thinking and writing about matters of nuclear deterrence strategy in terms of Game Theory or Rational Choice Theory, this second desideratum was a matter of defining the nation's utility function in the face of a set of hypothetical choices. But at this point, some of those intellectuals realized that they faced a very puzzling problem. To put it simply, should they find ways to build into the physical system and set of standing orders the preference structure that the relevant decision makers have now, or the preference structure they might have after the attack? After all, contemplating these end-times scenarios quietly in a backroom of the Pentagon, the planners might conclude that should America suffer the sort of devastating attack that would effectively terminate the existence of the United States as a functioning political entity, it would make no sense at all to launch a counter-attack whose sole purpose was the vengeful killing of several hundred million Soviet citizens, none of whom had played any role in the launch of the attack. But the defense intellectuals could also see that after the attack, with America in ruins, those still in control of nuclear weapons might desperately want revenge simply for the sake of revenge. In short, the trauma of the attack might change the preference order, or utility function, of the surviving decision makers.

Since the planners could recognize this possibility in advance, in accordance with which utility function should the plans be made? The one the decision makers had now, or the one they thought they were likely to have then?

If we step back from the horror of these speculations, we can see that this dramatic example is an instance of a much larger theoretically intractable problem. Rational Choice Theory assumes that utility functions are both exogenously given and invariant. The utility functions are exoge-

nously given in the sense that whatever determines them is outside of, or exogenous to, the system of decision being analyzed. The utility functions are invariant because, for purposes of the expected utility calculations, they are assumed to remain unchanged and are the foundation on which the calculations are based. So in situations in which the utility functions themselves change, the theory has nothing to say.

The same point can be made in another and more striking way. We have already seen that interpersonal comparisons of utility are not allowed in the theory of rational choice. The utility functions are cardinal, which is to say invariant under linear transformations, which in turn means that neither the units nor the zero point of two distinct utility functions are comparable. All of modern economic theory is erected on this assumption, by the way. [See the classic work by Lionel Robbins, *An Essay on the Nature and Significance of Economic Science*.] Now, from the point of view of Rational Choice Theory, a person simply is an embodied utility function. If a person's utility function changes, then so far as the theory is concerned, that person is now a new person, no longer the old person, and there can be no useful comparison of that person's utility function before and after the change, because that is the same as trying to compare the utility functions of two different people. In other words, the question posed by the defense intellectuals has no answer.

> 11 When I was a child, I spake as a child, I understood as a child, I thought as a child: but when I became a man, I put away childish things.
>
> 12 For now we see through a glass, darkly; but then face to face: now I know in part; but then shall I know even as also I am known. [1 *Corinthians* 13]

Now, if you think about it for even a moment, you will see that growing up, maturing, and aging is a process, common to all human beings, that among other things involves a change of one's utility function. Surely any useful theory of rational choice must allow for growth and change. But the Theory of Rational Choice does not, and cannot. That does, to put it mildly, seem to be a bit of a problem.

Well, so much for the Theory of Rational Choice, for the moment.[1]

We come now to the most elaborate, the technically most difficult, the most popular, and the most often misunderstood body of formal materials applied to philosophy, politics, law, military strategy, economics, and love: Game Theory. Quick check—Google finds 600,000 sites for "Prisoner's Dilemma" and 1,900, 000 sites for "zero sum game." Not quite up there with Lady Gaga [83,700,000], but still, not chopped chicken liver either. This is going to take a long time, and there are going to be some seriously technical patches that will try both your patience and my skills at explication. Nevertheless, if, in the immortal words of W. S. Gilbert in *Patience*, you want to "get up all the germs/ of the transcendental terms," now is your chance to do it. By the way, if you are actually paying attention to the outline with which I started, you will notice that I moved Game Theory up ahead of Collective Choice Theory. Arrow's Theorem (Arrow, 1963) and Amartya Sen's extension of it are two of the loveliest bits of theoretical material around, fully deserving of the two Nobel Prizes they earned. But their application to the fields you folks come from is not as rich as the application or misapplication of Game Theory, so I figured I would get to the good stuff before you all drift away. Here we go.

[1] *An explanatory word to my readers.* Each of the installments of this Formal Methods tutorial is not very long. There are two reasons for this. First, it is difficult stuff, and I do not want to scare away readers for whom this is all new. Second, I am writing two blogs at the same time—this one and my Memoirs—and I am working flat out. Five typed pages or so of this formal material is all I can manage each time I post. So be patient.

Chapter 3

The elements of Game Theory

3.1 Introductory remarks

In the eighteenth and nineteenth centuries, the standard conception of the capitalist market was of a place inhabited by so many sellers and so many buyers that the actions of any one buyer or seller had a negligible effect on prices. One buyer, by shifting to a different supplier or choosing not to buy at all, could have no measurable impact on what came to be called aggregate demand, and the output of one factory or shop had as little impact on aggregate supply. The marketplace was, in this sense, opaque. One could not see through the hustle and bustle to the individual suppliers or buyers whose actions, intersecting with one another, were determining the structure of prices. The standard term for this situation is that everyone was a "price taker," and no one was a "price maker." This was always an idealization that did not quite fit the facts, of course. First of all, from the very beginning there had been producers who managed to control so large a part of the market for their goods that they could simply dictate the prices at which they sold. They were said to have a monopolistic position in the market. There were also buyers who exercised a monopoly—Kings and Princes who by main force made themselves the sole buyers for certain luxury or military goods and so could dictate the prices at which they bought.

By the beginning of the twentieth century, economists were theoriz-

ing about situations that fell somewhere between monopoly and perfect competition, situations in which a small number of producers dominated a market—three or four great steel producers, three auto manufacturers, and so forth. A sizeable literature grew up dealing with what was called Imperfect Competition. For example, in 1933, Joan Robinson, the doyenne of the Cambridge School, published *The Economics of Imperfect Competition*. The defining characteristic of imperfect competition is that it is a situation in which the opacity of the market lifts, and it becomes possible for a producer to know about, be aware of the individual behavior of, and thereby adjust its own behavior to, that of the other producers.

Beginning in the 1920s, John von Neumann, one of the genuinely great minds of the last hundred years and more, developed a powerful mathematical analysis of the decision making that is possible in the precisely delineated structure of a game as well as in the situation of imperfect competition that economists had been examining. Joining forces with Oskar Morgenstern, an economist, von Neumann elaborated his theories in one of the great books of the twentieth century, *The Theory of Games and Economic Behavior*, published in 1944 (Neumann and Morgenstern, 1944). If you are unfamiliar with von Neumann's career, it is worth your time to look him up. He had a unparalleled capacity to grasp the underlying formal structure of a wide variety of fields, and made contributions not only to mathematics and economics, but also to physics. He is also the person who came up with the idea of using a binary number system so that it could be modeled in an electrical circuit, thereby making possible the digital age. Suffice it to say that there are only two or three talents that I would give years off my life to possess, and that is one of them.

What makes games so interesting, from von Neumann's point of view, is that they are interactions in which the number of players, the outcomes or payoffs, and the permissible moves are all precisely defined by clearly stated rules. Games are thus, in a sense, models of economic transactions. In many games, each player knows who his or her fellow players are, how they value the outcomes, and what the moves are at any stage in the game that are available to each player.

3.2 The definition of a game—extensive form

I am going to start by analyzing a very simple game. This will give me a chance to define the key concepts we are going to be working with. Here is a game I invented for these purposes, called Take Away Matchsticks. There are two players, and a pile of four matchsticks lying on a table. Players take turns moving. Each move consists of taking away either one or two matchsticks. The last player to take away a matchstick loses. The loser has to give the winner a penny. The rules say that if it is your turn, you have to move. That's the whole game. Obviously, the first person to move loses every time, because either she takes away one matchstick or two. If she takes away one, that leaves three, and the other player takes two, forcing her to take the last one and lose. If the first player starts by taking away two, then the second player takes one, forcing the first player to take the last matchstick and lose. Everybody got it? Trust me, for present purposes, you don't want me to choose a more complicated game.

A game consists of a number of moves. The rules define the starting point, the number of players, whose move it is at each step in the game, what the available choices are for each player at each step, when the game ends, and what the result is for each player at each ending point. We are going to limit ourselves to games whose rules define a finite number of players and guarantee a finite number of moves. So, for example, if we are analyzing chess, we will include the rule that says that if a position is repeated three times, or if fifty moves are made by each player without a piece being taken or a pawn being advanced to the eighth rank, then the game is declared a draw.

In Taking Away Matchsticks, there are two players, whom we will call A and B. The starting point is the pile of four matchsticks. The player labeled A goes first [tough luck]. A move consists of taking away one matchstick or two. The rules of the game ensure that it can last at most four moves [if each player takes one matchstick each time—stupid, but legal]. There are two possible outcomes: either A pays B one cent, or B pays A one cent. There cannot be any draws.

Just as Economics doesn't care how anything is actually produced or consumed, so Game Theory doesn't really whether a game is played with matchsticks or chess pieces or a pile of chips and a deck or a bat and a ball. It only cares about moves and players and outcomes and such. So our entire little game can be fully represented by the following diagram, which for obvious reasons is called a Game Tree:

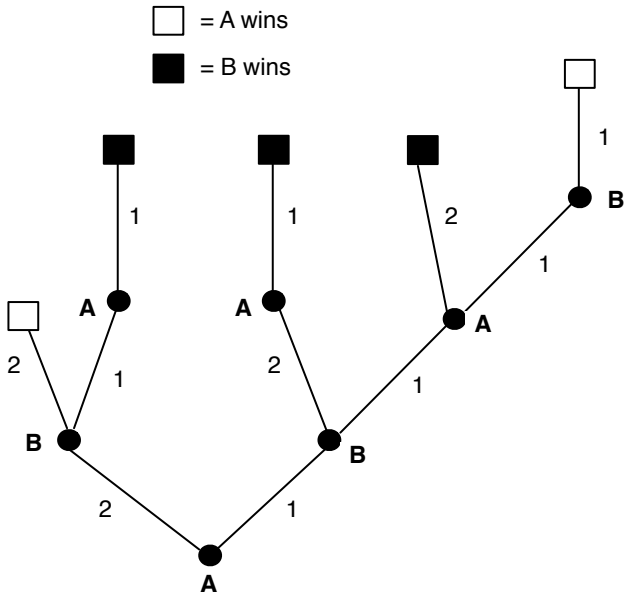


Figure 3.1: Game Tree

Each of the black circles is a node. It is a point at which a player has a move. The nodes are labeled, showing whose move it is. Each branch coming out of a node represents one of the moves that the rules allow that player to make at that point in the game. A square box indicates that the game is over. A white box indicates that Player A has won. A black box indicates that Player B has won. For purposes of Game Theory, two games with identical game trees are indistinguishable, even if one involves guns and swords and the other involves cards and chips.

If I had created a game in which a coin toss decides which player goes first [a little bit fairer for Player A], then there would be three players in

the game. The third player would be Lady Luck, and she would have the first move.

I want each of you to take a moment to look carefully at the Game Tree and follow out in your mind the sequence of moves I described earlier leading to one player or the other winning. There are five different sequences of moves that can occur in the game, each one leading to a different square box. Notice that there are no loops in the tree—no way that two different sequences of plays can lead to the same node up the tree a ways. For example: A takes the right hand branch; B then takes the right hand branch; then A takes the left hand branch. Game over, B wins. Or, A takes the left hand branch, B takes the right hand branch, A takes the only branch offered. Game over, B wins. And so forth. A game represented in this form is said to be in the Extensive Form. The Game Tree is said to be the Extensive Form of the Game.

Let us suppose A agrees to play the game with B and then is called away for an emergency. She asks the referee to play for her, following to the letter her instructions. [There has to be a referee to make sure no one cheats]. The referee agrees, but insists that A give him a complete set of instructions, so that no matter what B does, the referee will know how to play A's hand. A says: here is what I want you to do: Take 1 matchstick. If B takes 1, take 2. If B takes 2, take 1. This set of instructions is called a Strategy. It tells the referee what to do in every situation in which A has a choice. There is no need to specify what the referee is to do when A's move is forced by the rules. The referee is now totally prepared for all eventualities. How many strategies does A have, total, including the one she actually chose? Well, here they are: [A1] Take 1. If B takes 1, take 2. If B takes 2, take 1 [A2] Take 1. If B takes 1 take 1. If B takes 2, take 1. [A3] Take 2. Notice that strategy A3 is complete because once A takes 2, the rest of the game so far as she is concerned is forced. Now let us suppose B says, "Well, if A isn't going to be there, I will just leave my strategy choice with the referee also." What are B's strategies? [B1] If A takes 1, take 2. If A takes 2, take 2. [B2] If A takes 1, take 2. If A takes 2, take 1. [B3] If A takes 1, take 1. If A takes 2, take 1. [B4] If A takes 1, take 1. If A takes 2, take 2. So A has three strategies and B has four. There are

thus 3 x 4 = 12 possible pairs of strategy choices that A and B can leave with the referee. Notice [very important] that there is no communication between A and B. Each chooses a strategy by him or herself. Now, there are no chance elements in this game—no rolls of the dice, no spins of a wheel. Game Theory allows for that, but this game doesn't happen to have any such "moves by Lady Luck." Therefore, once you know the strategy choices of A and B, you can calculate the outcome of the game. And we know what the payoffs are. In each possible outcome, either A wins a penny and B loses a penny, or B wins a penny and A loses a penny. Notice that there is no assumption that a penny yields the same amount of utility to A as to B. Indeed, any such statement is meaningless. We are now ready to construct what Game Theory calls the "payoff matrix," which in this case is a grid three by four, each box of which represents the payoffs to A and B of a pair of strategies that are played against one another. For example, what happens if A tells the referee to play her first strategy, [A1], and B tells the referee [without knowing what A is doing] to play his first strategy, B1? Well, A1 tells the referee to take 1 stick. Then it is B's turn, and B1 tells the referee that if A takes 1, the referee is to take 2. Now it is A's turn, and she has no choice but to take the last matchstick. B wins, and A pays B one cent. So the payoff for A is -1, and the payoff for B is +1. Here is the complete payoff matrix for this game:

	B1	B2	B3	B4
A1	-1, +1	-1, +1	-1, +1	-1, +1
A2	-1, +1	-1, +1	-1, +1	-1, +1
A3	+1,-1	-1, +1	-1, +1	+1, -1

Table 3.1: Payoff matrix for a simple game.

You should take a few minutes to look at this carefully and be sure that you see how I derived the figures in the boxes—the payoffs. This is called the **normal form** of the game. If you look at the payoff matrix just above, you will see that B wins in all but two cases: when A plays strategy A3 and B plays either B1 or B4. Now, Game Theory assumes that both

players know everything we have just laid out about the game, so A and B both know the payoff matrix. B can see that if he chooses strategy B2 or B3, then he is guaranteed a win no matter what A does. Furthermore, either B2 or B3 is equally good for B. We describe this by saying that B2 and B3 are dominant strategies for B. A is out of luck. Her only hope, and a pretty slim one at that, is to play A3 and hope against hope that B is a dope. With this elementary example before us, let me now make several comments.

1. Legal theorists, political scientists, sociologists, philosophers all seem to think that there is something deep and profound about the Prisoner's Dilemma. Well, I invented the simplest game I could think of, and in that idiot game, there are three strategies for A and four for B. The Prisoner's Dilemma is a game with only two strategies for each player. How can something that much simpler than the idiot game I invented possibly tell us anything useful about the world? The truth is, it can't!

2. **From the point of view of Game Theory, the entire game is represented by the payoff matrix**. Any information not contained in the payoff matrix [like the fact that this game uses matchsticks, or that B has brown eyes] is irrelevant. All of the games with the same payoff matrix are, from the point of view of Game Theory, the same game. For a long time, until we get to something called Bargaining Theory, the little stories I tell about the games I am analyzing will serve simply to make the argument easier to follow. All the inferences will be based on the information in the payoff matrix. When we get to Bargaining Theory, which is tremendous fun but rather light on theorems, it will turn out that a great deal turns on what story you tell about the game. [For those of you who are interested, the classic work, which also won the author a Nobel Prize, is *The Strategy of Conflict* by Thomas Schelling (Schelling, 1960).]

3. In the game above, the only information we actually use about the payoffs is A's ordinal preference for the possible outcomes of the

game and B's ordinal preference for those outcomes. We make no use of the fact that the payoffs are money, nor do we use the fact that the amount of money won by one player happens to equals the amount of money lost by the other player. Even when we are talking about ordinal preference, I am going to use numbers, simply because they make it very easy to keep in mind the player's preference order.

4. At a certain point, when we introduce moves by Lady Luck [roll of the dice, spin of the wheel, etc.], we will have to shift up to cardinal preference orders for A and B. At that point, we will need cardinal numbers for the entries in the payoff matrices. The numbers before the comma will be A's utility for a certain outcome, as determined by A's cardinal utility function, and the numbers after the comma will be those for B. NOTHING AT ALL CAN BE INFERRED FROM THE NUMERICAL RELATIONSHIP BETWEEN AN ENTRY IN FRONT OF A COMMA AND THE ENTRY AFTER A COMMA. This is because the utility indices indicated by the numbers before the comma are invariant up to a linear transformation [or an affine transformation, as it is apparently now called, but I am too old to learn anything], and the same is true for the utility indices after the comma. If I multiply all of B's utilities for payoffs by one million, no information has been added or lost.

5. A is assumed to have a utility function that assigns ordinal [later cardinal] numbers to the outcomes. So is B. The outcomes are the terminations of the game as defined by the rules and diagrammed on the game tree. The rules may simply stipulate who is declared to have won and who has lost, or they may assign various payoffs, in money or anything else, to one or more of the players. No assumption is made about the attitudes of the players to these outcomes, save that their attitudes must generate consistent [ordinal or cardinal] preference orders of the outcomes. A can perfectly well prefer having B win the game over herself winning the game. Eventually, we will be assuming that both A and B are capable of carrying out expected utility calculations, and that each prefers an outcome with

a greater expected utility to one with a lesser expected utility. But that assumption does not have built into it any hidden assumptions about what floats A's boat. It is utility, not money or anything else, that A and B are maximizing.

6. We have been talking thus far only about two person games. The mathematical theory developed by von Neumann is capable of proving powerful theorems only for two person games. A great deal can be said about multi-person games, especially those allowing for pre-play communication, which leads to coalitions, betrayals, and all manner of interesting stuff. But unfortunately not much that is rigorous and susceptible of proof can be said about such games.

7. Really important: Game Theory treats the extensive form of a game [game tree] and the normal form of the game [payoff matrix] as equivalent. As we have already seen in the case of planning for nuclear war, that assumed equivalence can be problematic, because in the playing out of the game in extensive form, the utility functions of the players may change. We will talk some more about this later, but for now, we are going to accept Game Theory's treatment of the two forms of a game as equivalent.

Now we are ready to start.

I will progress from game to game [i.e., from payoff matrix to payoff matrix], making things a little more complicated each time, until we get to the payoff for all of this [so to speak]: a two person mixed strategy zero sum game. I will then explain [but probably not drive you nuts by actually proving] The Fundamental Theorem of Game Theory, von Neuman's own [if I may make a play on a popular line of environmentally friendly spaghetti sauces and such], which says that *Every two person zero-sum mixed strategy game has a solution in the strong sense.* But I get ahead of myself.

Here is our first little game. Two businesses are competing to produce and sell children's toys, and each one must decide whether to make hula hoops or yoyos [but not both, for technical reasons]. Each player knows everything there is to know about the costs, the market, and such, except

what the other player is going to choose to do. Although neither player knows what the other will do, each knows what effect the other's decision will have on the bottom line of each company. The following matrix shows the profit each will make in one of the four possible situations: A makes hula hoops and B makes hula hoops, A makes hula hoops and B makes yoyos, A makes yoyos and B makes hula hoops, or A makes yoyos and B makes yoyos. Now, I know that some of you are philosophers, but I really do not want you wasting your time wondering how they know these things, or what extraneous factors, cultural or otherwise, might affect their decisions. We will get nowhere if you do. Just go with the flow here. Table 3.2 is the payoff matrix for the game. We assume that both players prefer making more money to making less, and do not care about anything else.

	B1 make hula hoops	B2 make yoyos
A1 Make hula hoops	10000, 1000	8000, 6000
A2 Make yoyos	9000, 6000	6000, 5000

Table 3.2: Payoff matrix for the game of making hula hoops and yoyos.

A is in a position to make a rational decision, for consider: If A decides to make hula hoops, then no matter which choice B makes, A can be sure of doing better by sticking with hula hoops than by changing to yoyos. As the matrix shows, if B chooses $B1$, then $A1$ is better than $A2$ [because 10000 is bigger than 9000]. If B chooses B2, then A1 is better than $A2$ [because 8000 is bigger than 6000]. The strategy $A1$ is thus said to dominate the strategy $A2$. Let us from now on use the notation P_{ij} to mean the payoff to A for the strategy pair $[A_i, B_j]$, and the notation Q_{ij} to mean the payoff to B for the strategy pair $[A_i, B_j]$. We can se by looking at the matrix that $P11 > P21$ and $P12 > P22$. Notice several things about this game:

1. A does not need to know B's payoffs for the four possible combinations. In this very simple case, A's payoff schedule alone is enough to allow the calculation that $A1$ dominates $A2$.

2. As I have several times emphasized, we do not need to know the

actual dollar or utility payoffs to carry out this line or reasoning. All we need to know is that $P11 > P21$ and $P12 > P22$.

Now look at the situation from B's point of view. A quick look at the payoff matrix reveals that B cannot use the line of reasoning that solved the choice problem for A. B's preference pattern is $(Q12 = Q21) > Q22 > Q11$. Strategy B1 does not dominate strategy B2, because $Q12 > Q11$, and $B2$ does not dominate $B1$, because $Q12 > Q22$. Now B takes an important step forward. Since B knows that A is rational, and since B knows A's payoffs, B reasons that A will choose strategy A1, since it is A's dominant strategy. Knowing that, B now looks at the payoff matrix and sees that with A choosing strategy $A1$, B's best strategy is $B2$, because $6000 > 1000$. Thus, assuming that the figures in the payoff matrix are correct, that A is perfectly rational, and that A cares only about maximizing her utility as represented by the figures in the matrix, B can now solve the choice problem by choosing strategy $B2$, even though neither of the B strategies dominates the other. With this tiny step forward, we begin to develop a theory that takes account of and adjusts for the rational calculations of the other player. In this way, we move beyond the opacity of the marketplace, which is the central point of Game Theory. The games in which one or the other player has a strictly dominant strategy are relatively rare [and uninteresting]. When neither player has a strictly dominant strategy, we must extend our notion of what constitutes a "rational" choice. Consider the following game with a 3 x 3 payoff matrix (see Table 3.3). Once again, the preferences are ordinal, and I use numbers simply because it is easy to tell in what order A or B ranks two payoffs.

	B1	B2	B3
A1	3, 9	5, 6	1, 4
A2	0, 11	6, 2	2, 11
A3	4, 5	5, 19	3, 8

Table 3.3: 3 x 3 payoff matrix for a game.

A little careful examination reveals that neither A nor B has a dominant strategy. For example, strategy $A2$ is not dominant [i.e., does best

no matter what B does], because if B chooses $B1$, A would do better with either $A1$ or $A3$. Strategy $B3$ is not dominant for B because if A were to choose $A1$, B would be better off with $B2$ or $B1$. And so forth. What are A and B to do in this situation? This is, notice, exactly what we meant by "choice under uncertainty.' A and b know everything about the payoff matrix, therefore everything about possible outcomes and consequences of pairs of strategic choices, but neither of them can be certain what the other will do. In this situation, there are obviously a great many different ways A and B might proceed [leaving to one side for the moment engaging in industrial espionage to find out what the other is thinking.] If they had cardinal utility functions, and hence could say not merely in which order they prefer the possible outcomes but also by how much they do, they might decide to avoid any strategies that have any possibility at all of outcomes they consider really terrible. That might narrow their choices to the point where dominance considerations could be invoked. By the same token, they might decide to give preference to strategies that have at least some possibility of outcomes they prize very, very highly. And of course one might do one and the other the other. With cardinal preference orders, they might even cast about for some way to engage in expected utility calculations, although without any knowledge of the probabilities, that would be very difficult. At this point, von Neuman and Morgenstern propose an extremely conservative rule for choosing from among the available strategies. They suggest that the players, in effect, try to minimize the damage they may suffer as a consequence of the choices of their opponents. Let us, they say, define something we will call the "security level" of a strategy. The security level is the worst outcome that the strategy can produce. For example, look at strategy A1. Depending on what B chooses, A will get 3, 5, or 1. So her security level for A1, the worst she can do if that is her choice, is 1. By the same reasoning, we can see that the security level of $A2$ is 0, and of $A3$ is 3. Correspondingly, B's security levels for his three strategies are 5 for $B1$, 2 for $B2$, and 4 for $B3$. If the players adopt von Neuman and Morgenstern's rule for rational choice, then A will choose $A3$, which has the maximum security level of her three strategies, and B will adopt $B1$ for the same reason. The outcome of the

game will be the intersection of strategies $A3$ and $B1$, which, as we can see from the payoff matrix gives A 4 and B 5. This pair of strategy choices guarantees at least 4 to A and at least 5 to B. But notice the following fact: There are two other pairs of strategy choices that are better for both A and B. ($A1B2$) gives the 5 and 6, instead of 4 and 5, and ($A3B2$) gives them 5 and 19 instead of 4 and 5. Now, both A and B can see this, of course, but without what is called pre-play communication or any way of making commitments to one another, they have no way of reaching either of those mutually better [i.e. Pareto Preferred] outcomes. The problem is that if B chooses $B2$, A may switch to $A2$ to get 6, leaving B with 2, which is worse than he can guarantee to himself by following the von Neuman Morgenstern rule.

There are lots of games like this—or, to put it another way, lots of situations which, when analyzed as games, produce payoff matrices with these characteristics. We shall return to them later. [The Prisoner's Dilemma is one example]. Instead, let us take the next step forward in the evolution of the theory. Very often, in a game A and B have what are called "strictly opposed" preferences over the outcomes. What that means is simply that if you take all the possible outcomes of the game and rank them from best to worst from A's point of view, B will rank them in exactly the opposite order.

Somewhat more formally, where P_{ij} and Q_{ij} are the payoffs to A and B when A's strategy A_i a is played against B's strategy B_j, and P_{qr} and Q_{qr} are the payoffs to A and B when A's strategy q is played against B's strategy B_r, then:

(i) $P_{ij} > P_{qr}$ if and only if $Q_{qr} > Q_{ij}$ and

(ii) $P_{ij} = P_{qr}$ if and only if $Q_{qr} = Q_{ij}$

You might think that most of life is like this, and especially that all bargaining is, but a little reflection will convince you that that is not so. Think of the situation in which Jones has a house to sell and Smith wants to buy a house. They enter into a negotiation, which we can call a bargaining game. Suppose the lowest price for which Jones will sell the house is

$350,000 and the highest price Smith will pay for the house is $375,000. Jones cannot get more than $375,000 for the house, and Smith cannot get the house for less than $350,000. Clearly, within that $25,000 spread, they have strictly opposed preferences. But both of them have an interest in concluding a sale, rather than in having the bargaining break down because they cannot come to an agreement in that "bargaining space." So, simplifying considerably, if we suppose there are three possible outcomes, namely (1) a sale price of 355,000, (2) a sale price of 370,000, and (3) no sale price, then clearly Jones prefers (2) to (1) and (1) to (3). Smith prefers (1) to (2) and (2) to (3). Jones' preference order is 213 and Smith's is 123. These are NOT strictly opposed preference orders [because in both orders alternative 3 is last]. Thus, many real world situations to which we might want to apply Game Theory are not cases of strictly opposed preference orders. Now consider a simple example of strictly opposed preference orders. Suppose a married couple, Harry and John, are trying to decide where they will go for their vacation, and suppose that all either of them cares about is the weather. For Harry, the warmer the better; for John, the cooler the better. [So why did they get married?, you ask.] They play a game in which the outcomes are the different places they could go for the vacation. Obviously, if Harry prefers destination 1 to destination 2, because 1 is warmer than 2, then we can be sure that Harry will prefer destination 2 to destination 1. You get the idea. Table 3.4 shows the payoff matrix.

	B1	B2	B3	B4
A1	9, -9	-4, 4	2, -2	-1, 1
A2	-1, 1	3, -3	-1, 1	0, 0
A3	6, -6	4, -4	5, -5	3, -3
A4	-3, 3	5, -5	1, -1	-2, 2

Table 3.4: Payoff matrix for a game with 4 strategies per player.

This is a more difficult game to analyze, and not merely because each player has four strategies rather than two. The problem is that neither player has a strictly dominating strategy. Consider each of the eight strate-

gies in turn:

1. $A1$ is not best for A if B should choose $B2$, $B3$, or $B4$ [Because if $B2$ anything is better, if $B3$ $A3$ is better, if $B4$ $A2$ or $A3$ is better]
2. $A2$ is not best if B should choose $B1$, $B2$, $B3$, or $B4$
3. $A3$ is not best if B should choose $B1$, or $B2$
4. $A4$ is not best if B should choose $B1$, $B3$, or $B4$
5. $B1$ is not best if A should choose $A1$, $A2$, or $A3$
6. $B2$ is not best if A should choose $A2$, $A3$, or $A4$
7. $B3$ is not best if A should choose $A1$, $A3$, or $A4$
8. $B4$ is not best if A should choose $A1$, $A2$, or $A4$

Before we go on, make sure you understand how I arrived at this series of conclusions. Look just for a moment at strategy $B3$. B says to himself: "If A should choose $A1$, I will get -2 with $B3$. But I would get 9 with $B1$, 4 with $B2$, and 1 with $B4$. So clearly $B3$ does not do best for me no matter what A does, and that is what 'dominant strategy' means. So $B3$ is not a dominant strategy." The same reasoning leads A and b to conclude that neither one has a dominant strategy. Now let us adopt von Neuman and Morgenstern's proposal that the players seek to maximize their security levels. By the same process we followed a short while ago, we find that the security levels for the strategies available to A and B are:

$A1$ -4 $A2$ -1 $A3$ 3 $A4$ -3
$B1$ -9 $B2$ -5 $B3$ -5 $B4$ -3

So $A3$ and $B4$ are the strategies with the maximum security levels, and following von Neuman and Morgenstern's rule, the players choose the strategy pair $(A3B4)$ which, according to the payoff matrix yields the payoffs (3, -3). If A holds to $A3$, B cannot do any better by switching strategies, because the other payoffs to B in that row are -6, -4, and -5. If B

holds to strategy $B4$, A cannot any better by switching strategies, because the other payoffs to A in that column are -1, 0, and -2. A pair of strategies with this property is called an equilibrium pair of strategies.

The following fact is crucial: A pair of strategies (A_i, B_j) is in equilibrium if and only if the entry A_{ij} is the minimum of its row, A_i, and the maximum of its column, B_j.

Here is a proof of that important proposition:
To say that Ai and Bj are in equilibrium is to say that neither player can improve his or her payoff by a strategy switch so long as the other player holds firm. This means that A's payoff, A_{ij}, is larger than any other payoff in its column, these being the payoffs available to A when B is holding to the strategy B_j. By the same reasoning, B_{ij} is the largest, or most preferred, payoff to B in row A_i, since those are the payoffs available to B so long as A holds fast to A_i. But by hypothesis, A and B have strictly opposed preferences [this is where that crucial assumption comes in], so the outcome at A_{ij} will be the least preferred of all the payoffs in row A_i from A's point of view. Thus, if A_i and B_j are in equilibrium, it follows that A_{ij} will be the maximum of its column and the minimum of its row.

Conversely, suppose that A_{ij} is the maximum of its column and the minimum of units row. Since it is the maximum of its column, A can only do worse by switching to a different strategy so long as B holds fast. And since A and B have strictly opposed preferences, payoff B_{ij} must be the most preferred of its row, for A_{ij} is the least preferred of its row. So B can only lose by switching so long as A holds fast. But this is the definition of an equilibrium pair of strategies. Q. E. D.

We have wandered pretty far into the weeds here, so you should take a moment to go over this argument and make sure you understand it. It is a typical Game Theory argument, and you need to become comfortable with that way of reasoning. Remember, we have already talked about whether identifying security levels and choosing a strategy to maximize your security level is a rational way of proceeding in game that presents you with choice under uncertainty. It is interesting to note, as we will see much later, that Rawls adopts this notion of rationality in *A Theory of Justice* (Rawls, 1971), where he dramatizes it by saying, in effect [not a

quote], Design an institution as though your worst enemy was going to assign you a place in it. In such a case, pretty clearly, maximizing the payoff to the least favored role in the institution makes a good deal of sense.

Thus far, we have looked at games in which each player's maximum security level shows up in only one of the available strategies, but obviously there might be several strategies with identical security levels, and that security level could perfectly well be the maximum one. In that case, the rule to choose the strategy with the maximum security level does not tell player which strategy to choose. All of the strategies exhibiting the maximum security level are equally good, as far ass the rule is concerned. But if we cannot specify which strategy a player will choose, following the rule, then how can we know what the outcome of the game will be? Fortunately, when players have strictly opposed preferences, it makes no difference. The following is a very important fact: If strategy pairs (A_i, B_j) and (A_p, B_r) are both equilibrium pairs of strategies, then so too are the pairs (A_i, B_r) and (A_p, B_j). What is more, in that case $A_{ij} = A_{ir} = A_{pj} = A_{pr} and B_{ij} = B_{ir} = B_{pr} = B_{ij}$. So, no matter how A and b mix and match their strategies with the maximum security levels, the results will be the same. [Note: When I say this, I mean the payoffs to the players will be the same. The actual play of the game may differ according to which strategies A and B choose, but they don't care about that, by hypothesis. They only care about the payoffs. keep that in mind, because down the line, it could be problematic.] Here is the proof. Let us suppose that we have a payoff matrix that is n rows by m columns, or n x m. I am going to show you a central part of the total matrix that is large enough to include all four payoff pairs: (A_{ij}, B_{ij}), (A_{ir}, B_{ir}), (A_{pj}, B_{pj}), and (A_{pr}, B_{pr}).

1. $A_{pr} \geq A_{ir}$ because (A_{pr}, B_{pr}) is an equilibrium point, and hence A_{pr} is the maximum of its column. [Notice, the symbol $\geq$ means "equal to or greater than." That is the way my word processing program writes it.]

2. $A_{ir} \geq A_{ij}$ because (A_{ij}, B_{ij}) is an equilibrium point, and hence A_{ij} is

	...	B_j	...	...	...	B_r	...	...	...
...									
A_i		A_{ij}, B_{ij}				A_{ir}, B_{ir}			
...									
A_p		A_{pj}, B_{pj}				A_{pr}, B_{pr}			
...									
...									

Table 3.5: Central part of a large payoff matrix.

a minimum of its row.

3. $A_{ij} \geq A_{pj}$, same reasoning as (1)

4. $A_{pj} \geq A_{pr}$, same reasoning as (2). Hence

5. $A_{pr} \geq A_{ir} \geq A_{ij} \geq A_{pj} \geq A_{pr}$ Therefore

6. $A_{pr} = A_{ir} = A_{ij} = A_{pj}$

The same reasoning establishes that $B_{ij} = B_{ir} = B_{pr} = B_{pj}$ and therefore obviously (A_{ir}, B_{ir}) and (A_{pj}, B_{pj}) are also equilibrium pairs.

Just to review, the key to the proof is the fact that A and B have strictly opposed preference orders. If that is not the case, the argument clearly does not go through.

Everything we have said thus far assumes only ordinal preferences, but that is not going to be enough to allow us to analyze games that involve chance elements, or what I have somewhat facetiously been calling moves by Lady Luck [think Marlon Brando singing "Luck be a lady tonight" in the movie version of Guys and Dolls.] Suppose that at some point in a game the rules call for a roll of the dice, a flip of a coin, or a spin of a wheel, with some player's options determined by the outcome of the chance event. That is going to create problems for our analysis.

Here is the simplest game I could think of to illustrate this idea. A moves first, and she has a choice. She can choose not to toss a coin, in which case B has to choose between a move that has the payoff (2.4, -2.4) to A and B, and a move that has the payoff (2, -2) to A and B. Pretty

obviously, B will choose the latter. OR A can opt to flip a coin. If the coin comes up heads, the game ends with the payoff (1, -1). if the coin comes up tails, the game ends with the payoff (6, -6). Not much of a game, but it will do.

What should A do? If she opts not to toss the coin, she has a sure thing payoff [given that B is rational] of 2. if she opts to flip the coin, she has a one-half chance of a payoff of 1 and a one-half chance of a payoff of 6. Now, if we forget that the numbers in the example are ordinal labels, we might be tempted to suppose that A can solve her problem by engaging in an expected utility calculation. After all, (1/2 x 1) + (1/2 x 6) = .5 + 3 = 3.5 so A should apparently choose to flip the coin. But these are ordinals, not cardinals, and all we really know is that for A, the payoffs are ranked 6 first, 2.4 second, 2 third, and 1 fourth. This ranking is preserved if we re-label the 6 as a 2.5 That still makes it first, which is all the information we have. But now, when we carry out an expected utility calculation on the game, we have 2 versus (1/2 x 1) + (1/2 x 2.5) = 1.75. With these numbers, A should change her strategy choice.

Obviously, we cannot analyze games with chance elements unless we assume that the players have cardinal utility functions with utility assignments that are invariant under an affine [linear] transformation. Therefore, we need now to introduce the formal machinery required to allow us to talk about cardinal utility functions. This is going to get seriously gnarly, I am afraid. The faint of heart may wish to take a vacation for a day or two while I lay all of this out. I choose to go into this for two reasons: First, as my son Tobias, who is following this blog, pointed out to me at dinner several evenings ago, I am really a rather nerdy wonk when it comes to this stuff. I just plain like it. I hadn't realized that, but of course he is right. I think it is nifty. Second, one of the central ideological messages of this blog is that too many intellectuals and academics adopt the jargon and the style of argument of Game Theory without any real grasp of the assumptions that are embedded in what they are saying. They are like party crashers at a Mass who think that the Eucharist is just a light snack, oblivious to its theological meaning. For those who can handle it, I want to take you through the formal unfolding of the concept of a cardinal util-

ity function. For those who cannot handle it, I want to shock and awe you so that in the future, when you idly assume that someone has a cardinal utility function, you will at least know what lies beneath the surface of that assumption. Here goes.

We must begin with the notion of a probability distribution over a set of outcomes. Remember that all of this theory assumes that in a game there are a finite number of possible outcomes [maybe just win or lose, but maybe also lots of different money payouts, or even things like a trip to the zoo, a fur coat, a dinner date with Kevin Bacon, etc.] The convention in probability theory is that probabilities range from 1 to 0 inclusive. If a possible outcome has a probability of 1, that means it is certain to happen. If it has a probability of 0, that means it is certain not to happen. Thus, probabilities are expressed as real numbers between 1 and 0. Most of the time, they are expressed as a decimal point followed by some real number, like .4 or .125, and so forth.

We also assume that the possible outcomes are independent of one another, not part of or nested inside one another. So we cannot have two outcomes one of which is "I lose the game" and another of which is "I lose the game and have to pay fifty dollars to the person who won." If all of this is so, then the probability that either outcome O will happen or outcome P will happen is equal to the probability that O will happen plus the probability that P will happen. So, if the probability of O is .3 and the probability of P is .6, then the probability of either O or P is (.3 + .6) = .9. A little reflection will tell you that if there are three possible outcomes, O, P, and Q, and if it is certain that one or another of them will happen, then the sum of their probabilities must be 1. In other words, "O or P or Q" is certain to happen. A probability distribution over the set of possible outcomes is a set of numbers, each of which is between 1 and 0 inclusive [some of the outcomes may be certain not to happen, in which case they have probability zero] and all of which add up to 1. Another way to express this [hang on, I am really reaching with my word processing program here] is this: If the probability of outcome i is p_i then for all n possible outcomes from 1 to n, $\sum p_i = 1$. [Cripes, That wasn't worth the effort. Oh well.]

Following Luce and Raiffa, I am going to use the term "Lottery" to refer to an experiment that has built into it a probability distribution over the set of n possible outcomes in a game. For example, if you want to construct a lottery that has built into it a .5 chance of outcome O, a .25 chance of outcome P, and a 2.5 chance of outcome Q, you can make a wooden wheel with an arrow fixed to its center. You can then draw radii on the wheel dividing it into a half segment and two quarter segments. Then you can spin the arrow in such a way that there was an equal chance of the point of the arrow coming to rest anywhere on the wheel [remember to oil the bearings]. That wheel would be a lottery with the desired probabilities.

O.K. We already know that each player has a consistent ordinal preference over the set of possible outcomes of the game. But we also know that that information all by itself is not enough to authorize us to represent that preference order by cardinal numbers. We can certainly use cardinal numbers in payoff matrices to represent a player's preferences—that is what I have been doing. But we cannot treat them as cardinal numbers. We can only treat them as ordinals. Obviously we need more information about the player's preference structure if we are to define a cardinal preference order for him or her over the possible outcomes.

Before we state the six axioms that von Neuman and Morgenstern proved are sufficient to allow us to impute a cardinal utility function to a player, we need some more definitions and some more notation. [I warned you this would get gnarly.] First of all, we must extend our notion of a Lottery to something called a Compound Lottery. A Simple Lottery is a probability distribution over as set of outcomes, O,P, Q, etc. A Compound Lottery is a Lottery the prizes in which are tickets in other lotteries over O, P, Q, etc. To make this as clear as I can, let me take a very simple case. Imagine a set of three outcomes (O, P, Q):

O = +$5

P = +$8

Q = -$10 [i.e., the player has to play ten dollars]

and a Lottery, L1, with three prizes, namely tickets in Lotteries L11, L12, and L13. L1 is set up so that there is a:

a .5 chance of winning a ticket in L11,
a .25 chance of winning a ticket in L12, and
a .25 chance of winning a ticket in L13.

The prizes in the Lotteries L11, L12, and L13 are the outcomes (O, P, and Q) Now, L11, L12, and L13 are probability distributions over O, P, Q, etc. So, let us suppose that these three Lotteries are in fact:
L11: .3 chance of O, .4 chance of P, and .3 chance of Q
L12: 1. chance of O, .0 chance of P, .9 chance of Q
L13: .8 chance of O, .1 chance of P, .1 chance of Q

Notice that in each case the probabilities add up to 1.

If our player, A, buys a ticket in $L1$, what is her chance of ending up with each of the outcomes, O, P, or Q? Well, she has a .5 chance of winning a ticket in L11, and L11 in turn offers a .3 chance of O, so that gives A so far a (.5)(.3) = .15 chance of O.

She also has a .25 chance of a ticket in L12, and L12 offers a .1 chance of O, so that gives her a (.25)(.1) = .025 chance of O.

Finally, she has a .25 chance of a ticket in L13, which offers a .8 chance of O, so that gives her a (.25)(.8) = .2 chance of O.

Adding .15, .025, and .2, we get a .375 chance of O.

If you carry out the same calculations for outcomes P and Q, you will find that A has a .225 chance of getting P and a .4 chance of getting Q, and sure enough, .375 + .225 + .4 = 1.

Now, what is the money value to A of this gamble? It is the Mathematical Expectation, or: (.375)(5) + (.225)(8) + (.4)(-10) = 1.875 + 1.8 - 4 = -.325 or minus 32.5 cents.

So, if all A cares about is making money, she ought not to buy a lottery ticket in L at any price. In fact, she should not even play if someone gives her a ticket.

This is what is called reducing a Compound Lottery to a Simple Lottery, and it should be obvious that you can do this with any finite number of prizes and any number of levels of lotteries of lotteries of lotteries. Notice one small point that will be important later: If a Lottery

offers a chance of p for one outcome, O, and chances of zero for all the other outcomes save a single one, Q, then the probability for Q will be $(1 - p)$. At long last, we are ready to state the six assumptions about someone's preferences, or Axioms, as von Neuman and Morgenstern call them, the positing of which is sufficient to allow us to deduce that the person's preferences over a set of outcomes can be represented by a Cardinal Utility Function. There is a very great deal of hairy detail that I am going to skip over, for two reasons. The first is that I want there to be someone still reading this when I get done. The second is that it is just too much trouble to try to get all this symbolism onto my blog. You can find the detail in Luce and Raiffa. O.K., here we go.

Assume there is a set of n outcomes, or prizes, $O = (O_1, O_2, \ldots, O_n)$

Axiom 1 *The individual has a weak preference ordering over O, with O_1 the most preferred and O_n the least preferred, and this ordering is complete and transitive. Thus, for any O_i and O_j, either O_i R O_j or O_j R O_i. Also, If O_i R O_j, and O_j R O_k, then O_i R O_k.*

[I told you we would use that stuff at the beginning.]

Axiom 2 (A biggie) *The individual is indifferent between any Compound Lottery and the Simple Lottery over O derived from the Compound Lottery by the ordinary mathematical process of reducing a compound lottery to a simple lottery [as I did for the example].*

This a very powerful axiom, and we have already met something like it in our discussion. In effect, it says that the individual has neither a taste for nor an aversion to any distribution of risk. The point is that the Compound Lotteries may exhibit a very broad spread of risk, whereas the Simple Lottery derived from them by the reduction process may have a very narrow spread of risk. Or vice versa. The individual doesn't care about that.

Axiom 3 *For any prize or outcome O_i, there is some Lottery over just the most and least preferred outcomes such that the individual is indifferent between that Lottery and the outcome O_i. A Lottery over just the most and least preferred*

outcomes is a Lottery that assigns some probability p to the most preferred outcome, O_1, and a probability $(1-p)$ to the least preferred outcome, O_n, and zero probability to all the other outcomes. Think of this as a needle on a scale marked 0 to 1. You show the person the outcome O_i, and then you slide the needle back and forth between the 1, which is labeled O_1 and the 0 [zero] which is labeled O_n. Somewhere between those two extremes, this Axiom says, there is a balancing point of probabilities that the person considers exactly as good as the certainty of O_i. Call that point U_i. It is the point that assigns a probability of U_i to O_1 and a probability of $(1-U_i)$ to O_n.

We are now going to give a name to the Lottery we are discussing, namely the Lottery $[U_iO_1, (1-U_i)O_n]$. We are going to call it $\tilde{O}_i$. Thus, according to this Axiom and our symbolism, the player A is indifferent between O_i and $\tilde{O}_i$.

If you have good mathematical intuition and are following this closely, it may occur to you that this number between 1 and 0, U_i, is going to turn out to be the Utility Index assigned to O_i in A's cardinal utility function. You would be right.

This Axiom is essentially a continuity axiom, and it is very, very powerful. It implies a number of important things. First, it implies that A does NOT have a lexicographic preference order. All of the outcomes are, in A's eyes, commensurable with one another, in the sense that for each of them, A is indifferent between it and some mix or other of the most and the least preferred outcomes. It also implies that we can, so far as A's preferences are concerned, reduce any Lottery, however complex, to some Simple Lottery over just O_1 and O_n. The Axiom guarantees that there is such a Lottery. Notice also that this Axiom implies that A is capable of making infinitely fine discriminations of preference between Lotteries. In short, this is one of those idealizing or simplifying assumptions [like continuous production functions] that economists make so that they can use fancy math.

Axiom 4 *In any lottery, $\tilde{O}$ can be substituted for O_i. Remember, Axiom 3 says that A is indifferent between $\tilde{O}_i$ and O_i. This axiom says that when you substitute $\tilde{O}_i$ for O_i in a lottery, A is indifferent between the old lottery and the new one.*

In effect, this says that the surrounding or context in which you carry out the substitution makes no difference to A. For example, the first lottery might assign a probability of .4 to the outcome O_i, while the new lottery assigns the same probability, .4, to $\tilde{O}_i$. [If you are starting to get lost, remember that $\tilde{O}_i$ is the lottery over just O_1 and O_n, such that A is indifferent between that lottery and the pure outcome O_i.]

Axiom 5 *Preference and Indifference among lottery tickets are transitive relations. So if A prefers Lottery 1 to Lottery 2, and Lottery 2 to Lottery 3, then A will prefer Lottery 1 to Lottery 3. Also, if A is indifferent between Lottery 1 and Lottery 2, and is indifferent between Lottery 2 and Lottery 3, then A will be indifferent between Lottery 1 and Lottery 3. This is a much stronger Axiom than it looks, as we shall see presently.*

If you put Axioms 1 through 5 together, they imply something very powerful, namely that for any Lottery, L, there is a lottery over just O_1 and O_n, such that A is indifferent between L and that lottery over O_1 and O_n. We need to go through the proof of this in order to prepare for the wrap up last axiom.

Let L be the lottery $(p_1O_1, p_2O_2, \ldots, p_nO_n)$.

Now, for each O_i in L, substitute $\tilde{O}_i$. Axioms 3 and 4 say this can be done. So, using our previous notation, where xIy means A is indifferent between x and y,

$(p_1O_1, \ldots, p_nO_n)$ I $(p_{11}, \ldots, p_{nn})$ so, expanding the right hand side,

$(p_1O_1, \ldots, p_nO_n)$ I $(p_1[U_1O_1, (1-U_1)O_n]), \ldots, (p_n[U_nO_n, (1-U_n)O_n)$ or, multiplying

$(p_1O_1, \ldots, p_nO_n)$ I $([p_1U_1 + p_2U_2 + \ldots + p_nU_n]O_1, [p_1\{1-U_1\} + \ldots + p_n\{1-U_n\}O_n])$ or

$(p_1O_1, \ldots, p_nO_n)$ I $([p_1U_1 + p_2U_2 + \ldots + p_nU_n]O_1, [p_1\{1-U_1\} + ldots + p_n1-U_n]O_n)$

if we let $p = p_1U_1 + p_2U_2 + \ldots p_nU_n$ then we have:

$(p_1O_1, \ldots, p_nO_n)$ I $(pO_1, (1-p)O_n)$. In other words, the lottery, L, with which we started is indifferent to a lottery just over the best and worst outcomes, O_1 and O_n.

Axiom 6 *The last axiom says that if p and p' are two probabilities, i.e., two real numbers between 1 and 0, then: $(pO_1, [1-p]O_n)$ R $(p'O_1, [1-p']O_n)$ if and only if $p \geq p'$.*

This Axiom says that the individual [A in our little story] prefers [or is indifferent between] one lottery over the best and the worst alternatives to another lottery over those same two alternatives if and only if the probability assigned to O_1 in the first lottery is equal to or greater than the probability assigned to O_1 in the second lottery.

Now, let us draw a deep breath, step out of the weeds, and remember what we have just done. First, we started with a finite set of outcomes, $O = (O_1, O_2, \ldots, O_n)$. Then we defined a simple lottery over the set O as a probability distribution over the set O. Then we defined a compound lottery as a lottery whose prizes include tickets in simple lotteries. At this point, we introduced five AXIOMS or assumptions about the preferences that our sample individual A has over the set of outcomes and simple and compound lotteries of those outcomes. These are not deductions. They are assumptions. Then we showed that these five Axioms, taken together, imply a very powerful conclusion. Finally, we introduced a sixth Axiom or assumption about A's preferences.

That is where we are now. von Neuman now takes the last step, and shows that if someone's preferences obey all six Axioms, then that person's preferences can be represented by a cardinal utility function over those outcomes that is invariant up to an affine (linear) transformation. I am not going to go through the proof, which consists mostly of substituting and multiplying through and gathering terms and all that good stuff. Suffice it to say that when von Neuman gets all done, he has shown that one way of assigning utility indices to the outcomes in O in conformity with the six Axioms is to assign to each outcome O_i the number U_i [as defined above]. This is then "the utility to A of O_i." Remember that this is just one way of assigning A's utility indices to the outcomes in the set O. Any affine transformation of those assignments will serve just as well.

All of this has to be true about A's preferences in order for us to say that A's preferences can be represented by a cardinal utility function.

I want now to take some time to make sure that everyone understands just how strong these assumptions are, and also exactly how to interpret them. The first point to understand is in a way the hardest. You might think that our subject, A, decides how she feels about all of these simple and compound lotteries by carrying out expected utility calculations and then saying to herself, "Well, since this one has a greater mathematical expectation than that one, I prefer this one to that one." You might think that, because, good heavens, how else could she possibly decide which she prefers to which? But if you thought that [which of course none of you does], you would be WRONG, WRONG, WRONG! TOTALLY WRONG, WRONG, WRONG! That would be, to use correctly a phrase that these days is almost always used incorrectly, begging the question. It would be assuming what is to be proved, and thus arguing in a circle. What von Neuman actually supposes is that our subject, A, looks at the outcomes O_1, O_2, etc and decides how she feels about them. She ranks them in order of her preference. She then looks at the infinitude of simple lotteries and compound lotteries and decides how she feels about them as well. She merges this all in her mind into a single complete, transitive ordering of all of those outcomes and simple lotteries and compound lotteries. **Then** von Neuman posits that her preferences, thus arrived at, in fact obey the six Axioms. **If that is so**, then, von Neuman shows, her preferences can be represented **AS THOUGH** she were carrying out expected utility calculations in her head in accordance with the axioms.

We are talking here about an enormously powerful set of idealizing and simplifying assumptions, as powerful in their way as the assumptions economists have to make before they can talk about continuously twice differentiable production functions [which they need in order to prove their nifty equilibrium theorems.] Let me draw on something I said earlier to show you just how powerful these Axioms are.

Look at Axiom 5, the transitivity axiom, and let us recall the eye doctor example. Suppose that the lotteries A is comparing are big Amusement Park wheels, on which are marked off different sized wedges [each defined

by two radii], each one of which is associated with one of the outcomes in the set, O. It would be no problem at all to construct a whole series of wheels, each of which is such a tiny bit different from the one next to it that when A is shown the two wheels together, she looks at them and says, I am indifferent between those two lotteries." But suitably arranged, the series of wheels might very slowly, indiscernibly, alter the size of the wedges associated with two prizes or outcomes, O_i and O_j, until, if we were to show A the first and the last in the series, she would look at them and say, flatly, I prefer the one on the left to the one on the right. Whoops. No transitivity! Axiom 5 rules out any such state of affairs.

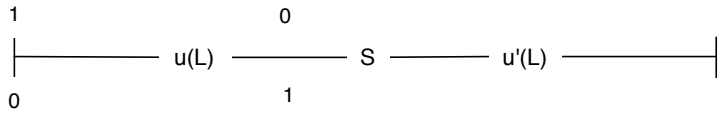
Well, you can think about each one of the Axioms and see whether you can imagine a situation in which the assumption of that Axiom clearly requires something very strong and even counterintuitive. But rather than go on about that, I am going to take the next step. We are now ready to extend our notion of strictly opposed preference orders. Recall that we describe the preference orders of A and B over a set of outcomes, O, as "strictly opposed" when A prefers O_i to O_j or is indifferent between them if and only if B prefers O_j to O_i or is indifferent between them. We will describe the preference orders of A and B over the infinite set of lotteries, simple and compound, over the set of outcomes, O, as "strictly competitive" when A prefers Lottery L_1 to Lottery L_2 or is indifferent between them if and only if B prefers L_2 to L_1 or is indifferent between them. This means that A and B not only rank all of the outcomes in exactly opposite ways. They also rank all of the lotteries, simple or compound, over those outcomes in exactly opposite ways.

In this very specific set of circumstances [where all six axioms apply to both A's preferences and B's preferences, and A and B have strictly competitive preferences], we can normalize the utility functions of A and B so that for any lottery, L, simple or compound, over the set of outcomes, O, the sum of the utility index assigned to L by A's utility function and the utility index assigned to L by B's utility function is a constant. This is what is meant by saying that a game played by A and B is a **constant sum game**.

Rather than grind out an algebraic proof, I will offer a simple, intu-

itive proof that should be easy to grasp. We shall use $u(L)$ to mean the utility that A's utility function assigns to L, and $u'(L)$ to mean the utility that B's utility function assigns to L. Now, we are permitted arbitrarily to let A's most preferred outcome, O_1, have a utility of 1, and A's least preferred outcome have a utility of 0. Since A and B have strictly opposed preferences for outcomes, B's most preferred outcome is O_n and his least preferred outcome is O_1. We are permitted to set B's utility for On equal to 1 and for O_1 equal to 0. So the utility assignments of both A and B can be portrayed as lying along a line that runs between 1 and 0.

No matter what lottery, L, we have chosen, we know from the Axioms that it is equivalent, for A, to some lottery over just O_1 and O_n whose probability weights are u and $(1-u)$ for some u. Think of that as a point somewhere on the line running between 1 and 0. [Remember that for the best and worst alternatives, O_1 and O_n, the point is an endpoint of the line.] The same thing is true for B. We are now going to prove that the point on the line representing A's utility for L and the point on the line representing B's utility for L are the same point. To prove this, we will assume the contrary and derive a contradiction with our assumption that A and B have strictly opposed preferences. So, let us choose a point representing $u(L)$ and a different point representing $u'(L)$, and then choose some point that lies between those two points, which we shall call S. Here is a picture of the situation. The line runs from 1 to 0 for A, and from 0 to 1 for B:



The point S represents a lottery, L_s, with weights S for O_n and $(1-S)$ for O_1. Now, just from looking at the diagram, we can see the following:

(i) A prefers L to L_s, because L puts greater weight on O_1 than L_s does. [$u(L)$ is closer to the 1 than S is].

(ii) B prefers L to L_s, because L puts greater weight on O_n [his favorite] than L_s does. [$u'(L)$ is closer to his 1 than S is.]

But this means that A and B do not have strictly opposed preferences, since they both prefer L to L_s. And this contradicts the assumption. So no matter which lottery L we choose, there cannot be a point S between $u(L)$ and $u'(L)$, which means they are the same point.

But if they are the same point, then A's utility is u and B's utility is $u' = (1 - u)$, regardless of which lottery, L, we choose. and:

$$u + u' = u + (1 - u) = 1$$

Now, B's utility function is invariant under an affine (linear) transformation. So let us introduce the following affine transformation:

$$u'' = u' - 1$$

What this does is to re-label B's utility assignments so that instead of running from 1 to 0, the run from 0 to -1. This means that A's and B's utilities for any arbitrary lottery L are no longer u and $(1 - u)$. Instead, they are now u and $-u$. And the sum of u and $-u$ is zero.

This, and only this, is what is meant by saying that a game played by A and B is a zero-sum game.

Now let us introduce the concept of a Mixed Strategy. All along, we have been working with two-person games whose rules allow for only a finite number of moves [with cut-off points like the rules that limit a chess game]. The Game Tree for such a game, however complex, defines a finite number of strategies for each player, even though that number may, as we have seen, be large even for very simple games. With a finite number of strategies for each player, we can convert the extensive form of the game to the normal form by constructing a payoff matrix with a finite number of rows representing A's strategies and a finite number of columns representing B's strategies. [From a mathematician's point of view, it doesn't matter how big the number of rows and columns, so long as they are finite in number]. But as we have seen, even with strictly opposed preferences, a game in which both players seek to maximize their security level may not have a stable equilibrium solution. Now von Neuman takes the final

step by introducing the concept of a mixed strategy.

When A gives her instructions to the Referee, she need not specify one of her pure strategies. Instead, she can define a probability distribution over her pure strategies and instruct the Referee to construct a Lottery that embodies that distribution. Just to be clear, this is a Lottery in which the "prizes" are strategies, not outcomes. Before leaving for her appointment, A tells the Referee to spin the wheel that has been constructed and play whichever strategy comes up. This is a real spin of the wheel. Neither A nor the referee can know which pure strategy will be played until the wheel has been spun. B can do the same thing, of course.

Each Lottery, or probability distribution over the set of pure strategies, is a mixed strategy, and quite obviously there are an infinite number of them. With an infinite number of mixed strategies for A, and an infinite number for B, there are of course also an infinite number of **mixed strategy pairs**, which is to say pairs each of which consists of one mixed strategy for A and one mixed strategy for B. Notice that a pure strategy now becomes simply a mixed strategy in which all the probability weights but one are zero.

For any mixed strategy pair, A can calculate the value to her of those mixed strategies being played against one another, although it is obviously tedious to do so. She says to herself: Suppose I play mixed strategy MA_1 and B plays mixed strategy MB_1. MA_1 offers a .3 probability of my playing pure strategy A_1, and MB_1 offers B a .2 probability of his playing pure strategy B_1, so I will calculate the payoff to me of A_1 played against B_1 and then discount that payoff, or multiply it, by (.4)(.1) = .04. Then I will do the same for A_2 against B_1, etc. Then I will add up all the bits of payoffs, and that is the value to me of the mixed strategy pair (MA_1, MB_1). I hope by now this is clear and reasonably obvious.

However, we can no longer construct a payoff matrix, because there are infinitely many mixed strategies for each player. Instead, we need a space of points that represent the payoffs to A of each of the infinite pairs of mixed strategies. Since we are dealing now with strictly competitive zero-sum games, we do not need to represent the payoff to B. Under the normalization we have chosen, that is simply 1 minus the payoff to A.

At this point I must do what mathematicians call "waving their hands." That is, instead of giving rigorous definitions and proofs [not that I have been doing that so far], I must simply wave my hands and describe informally what is going on, and hope that you have enough mathematical intuition to grasp what I am saying. To represent all of the mixed strategy pairs and their payoffs to A, we are going to need a space with $(n-1)+(m-1)+1$ dimensions.

The first $(n-1)$ dimensions will represent the probability weights being given to A's n pure strategies. [$(n-1)$ because the weights must add up to 1, so once you have specified $(n-1)$ of them, the last one is implicit.] The next $(m-1)$ dimensions represent the probability weights being given to B's m pure strategies. The last dimension, which you can think of intuitively as the height of a point off of a hyperplane, represents A's payoff. We only need to represent A's payoff because this is a zero-sum game, and B's payoff is just the negative of A's. Obviously, we are only interested in the part of the space that runs, on each axis, from 0 to 1, because both the probability weights and the payoffs all run from 0 to 1 inclusive.

It is said that the great Russian English mathematician Besicovitch could visualize objects in n-dimensional vector space. If he wanted to know whether something was true, he would "look" at the object in the space and rotate it, examining it until it was obvious to him what its properties were. Then he would take out pen and paper and go through the tedium of proving what he already knew was true. I suspect the same must have been true of von Neuman. Well, God knows it isn't true of me, so I must just soldier on, trying to connect the dots.

You and I can get some visual sense of what such a space would be like by thinking of the simplest case, in which A and B each have only two strategies. In that nice simple case, the number of dimensions we need is (2-1) + (2-1) + 1, or 3. And most of us can imagine a three-dimensional system of axes. Just think of a three dimensional graph with an x-axis and a y-axis forming a plane or bottom, and a z-axis sticking up. The infinity of A's mixed strategies can be represented as points along the x-axis running from the origin to plus 1. The origin represents the mixed strategy with zero weight given to A_1, and therefore a weight of 1 given

to A_2. In other words, it represents the pure strategy A_2. Any point along the line represents some mixture of A_1 and A_2. The point 1 on the x-axis represents the pure strategy A_1. Same thing on the y-axis for B's strategies B_1, B_2, and the mixtures of them. We thus have a square bounded by the points (0,0), (1,0), (0,1), and (1,1). The z-axis measures the payoff to A for each point in that square, and that set of points between 0 and 1 in height that together form a surface over the square.

Now [here goes the hand-waving], the function mapping mixed strategy pairs onto payoffs is a continuous one, because a tiny change in the assignment of probability weights results in a tiny change in the payoff. [I hope that is obvious. If it isn't, take my word for it—which is a great thing for a philosopher to say who is trying to explain some mathematics, I know, but I have my limits!]

OK. Got that in your mind's eye? Now, let us recall that von Neuman offered a "solution" of strictly competitive games in terms of something called security levels and equilibrium pairs of strategies. Suppose that somewhere in that space of payoffs, there is a point that represents the payoff to a pair of equilibrium mixed strategies. What would that mean and what would it look like?

Well, what it would mean is this: First of all, if B holds to his mixed-strategy choice, any movement A makes back and forth along the x-axis is going to be worse for her. [That is what it means to maximize your security level]. Visually, that means that as she moves back and forth along the x-axis, the point in space representing the payoff to her goes down. What is more, because of continuity, it goes down smoothly. A little movement one way or the other produces a little move down of the payoff point. A bigger move one way or the other produces a bigger move down. For B, the whole situation is reversed, because B's payoffs are equal to the negative of A's payoff. [Zero-sum game]. So, if A holds to her mixed strategy choice, any movement B makes along the y-axis will push the payoff point up [up for B is bad, because the payoff point is A's payoff, and B's is the negative of that.]

Now, what does this region of the payoff surface look like? Well, the point we are focusing on has the property that if you move back or forth

in one dimension the surface falls off, and if you move back or forth in the other dimension, the surface climbs. Have you ever watched a Western movie? Have you ever ridden a horse? Can you see that the surface looks like a saddle? Especially a Western saddle, which has a pommel to grab onto and a nice high back to the seat. The point right under you on the saddle, the point you are sitting on, is a "saddle point." It has the property that if you run your finger from that point side to side [along the y-axis], your finger goes down, and if you run your finger from that point front or back, your finger goes up.

Now we know what an equilibrium point looks like, at least in three dimensional space, for the case in which A and B each have two pure strategies. Exactly the analogous thing would be true of a hyperplane in hyperspace [you can get your light sabers at the desk before we go into warp speed]. So, we can say that **if there is a saddle point in the space representing a two person zero sum mixed strategy game, then that point will occur at the intersection of an equilibrium pair of mixed strategies, and in that sense will be a Solution to the game**. So, are there such points? Now comes the boffo ending for which all of this has been a preparation. John von Neuman proved the following theorem: **Every two person zero sum mixed strategy game has a solution. That solution is represented by a saddle point in the n-dimensional vector space representing the normal form of the game**. I really think von Neuman must have seen this in one exquisite flash of mathematical intuition, and then just cranked out a proof of a proposition he could just see is true. I am not going to go through the proof [I am not completely crazy], but having brought you this far, I think I owe it to you to just tell you the idea underlying it.

In a nutshell, here it is. Von Neuman defines a continuous transformation or mapping of the strategy space onto itself. He then proves that the transformation has this neat property: a point is mapped by the transformation onto itself [i.e., it remains invariant under the transformation] if and only if that point is a saddle point, and is thus a solution to the game. He then appeals to a famous theorem proved by the great Dutch mathematician L. E. J. Brower, which states that every continuous transfor-

mation of a compact space onto itself has at least one fixed point, which is to say a point that the transformation maps onto itself. [Hence, this is known as the Fixed Point Theorem.] Ta Da! [Can you believe that when I taught this stuff to my philosophy graduate students at UMass, I not only proved von Neuman's theorem, I even proved Brouwer's theorem? Ah, there were giants in the earth in those days, as the Good Book says.]

And that is it, folks. That is the high point of formal Game Theory. There is a vast amount more to say, and I am going to say a good deal of it, but the subject never again rises to this level of formal elegance or power. Notice, before we move on, one important fact. von Neuman proves that every zero-sum two person mixed strategy game has a solution. But since Brower's theorem just tells you there exists a fixed point, and doesn't tell you how to find it, in general Game Theory cannot tell us how to solve even this limited category of games. [If there is anyone out there who has ever been involved with Linear Programming, every Linear Programming problem is equivalent to a zero-sum mixed strategy two person game, so that is why, in a certain sense of why, you also cannot be sure of solving a Linear Programming problem.] Oh yes, one final point, which we have already encountered in a simpler form. if there are two saddle points, they are equivalent, in the sense that they give the same payoffs to A and to B.

Once again, let us pause to catch our breath. We arrived at this magnificent theorem by making a series of very powerful constraining and simplifying assumptions. Let us just list some of them:

0. We began by talking about games.

1. We limited ourselves to two person games

2. We limited ourselves to players whose preferences satisfy the six powerful Axioms from which we can deduce that their preferences can be represented by cardinal utility functions.

3. We limited ourselves to players with strictly competitive preferences

4. We allowed for mixed strategies.

5. We accepted mathematical expectation as a rational way of calculating the value of a strategy involving elements of risk.

6. We adopted von Neuman's extremely conservative rule of choice of strategies —maximizing the security levels.

7. We assumed no pre-play communication between the players.

8. We assumed perfect knowledge by both players of the information required to construct the payoff matrix or payoff space.

Every one of these assumptions can be altered or dropped. When that happens, a vast array of possibilities open up. No really powerful theorems can be proved about any of those possibilities, but lots and lots can be said. Here is how I am going to proceed. First, I am going to discuss each of these assumptions briefly and sketch the sorts of possibilities that open up when we drop it or alter it. After that, I will gather up everything we have learned and apply it to a number of specific texts in which Game Theory concepts are used. I will offer a discussion of the so-called Prisoner's Dilemma, a full scale analysis of John Rawls' central claim in *A Theory of Justice* (Rawls, 1971), a critique of Robert Nozick's *Anarchy, State, and Utopia* (Nozick, 1974), a detailed critique of a book by Jon Elster called *Making Sense of Marx* (Elster, 1985), a critique of the use made of Game Theory by nuclear deterrence strategists, and some remarks on the use of Game Theory concepts in writings by legal theorists. By then, you ought to be able to carry out this sort of critique yourselves whenever you encounter Game Theoretic or Rational Choice notions in your field of specialization.

Now let me say something about each of the nine assumptions listed above.

3.2.1 The Modeling of Real Situations as Games

I identify this as assumption zero because it is so fundamental to the entire intellectual enterprise that it is easy to forget what a powerful simplification and idealization it is. Games are activities **defined by** rules.

Imagine yourself watching two people playing chess, not knowing what chess is, but knowing only that a game is being played in the area. How would you describe what you are watching? Which of the things you see are appropriately included in the game and which are extraneous? Which characteristics of the various objects and people in the neighborhood are part of, or relevant to, the game? Is gender relevant? Is race relevant? Is the dog sitting by the table part of the game? Are the troubled sighs of one of the persons a part of the game? How do you know when the game begins and when it ends? Is the clothing of the persons in the area relevant? Are all of the people in the area part of the game, or only some of them? Indeed, are any of them part of the game? You cannot answer any of these questions easily without alluding to the rules of the game of chess. Once you acquaint yourself with the rules of chess, all of these questions have easy answers.

Now imagine yourself watching a war. Not one of the questions I raised in the previous paragraph has an obvious answer with regard to a war. When does a war start and when does it end? Are the economic activities taking place in the vicinity of the fighting part of the war or not? Who are the participants in a war? States that have formally declared war on one another, other nearby states, private individuals? And so forth. War is not a game. I don't mean that in the usual sense—that it is serious, that people get killed, etc. I mean it in the Game Theory sense. War is not an activity defined by a set of rules with reference to which those questions can be answered. Neither is market exchange, contrary to what you might imagine, nor is love, nor indeed is politics. There are many **descriptive** generalizations you can make about war, market exchange, love, and politics, but no statements that are **determinative** or **definitive** of those human activities. When you apply the concepts of Game Theory to any one of them, you are covertly importing into your discussion all the powerful simplifications and rule-governed stipulations that permit us to identify an activity as a game. Whenever you read an author who uses the concepts of Game Theory [move, payoff, strategy, zero sum, Prisoners' Dilemma, etc] in talking about some political or military or legal or economic situation, think about that.

3.2.2 Games with more than two persons

As soon as we open things up to allow for more than two players in a game, everything gets very complicated. First of all, with three or more players, no meaning can be given to the concept of opposed preference orders. We can still make the assumption of cardinal utility functions if we wish, because that is an assumption about an individual player's preference structure, and has no reference to any particular game. With three or more players, it also becomes difficult to represent the game by means of a payoff matrix. Not impossible—we can always define an n-dimension matrix—just very difficult either to visualize or to employ as a heuristic device for analyzing a game. That is why writers who invoke the concepts or the language of Game Theory will sometimes reduce a complex social situation to "a player and everyone else," in effect trying to turn a multi-player game into a two player game. That is almost always a bad idea, because in order to treat a group of people as one player, you must abstract from precisely the intra-party dynamics that you usually want to analyze.

Multi-player games also for the first time introduce the possibility of coalitions of players. Coalitions may either be overt and explicit, as when several players agree to work together, or they may be tacit, as when players who are not communicating overtly with one another begin to adjust their behavior to one another in reciprocal ways for cooperative ends. Once we allow for coalitions, we encounter the possibility of defections of one or more parties from a coalition, and that leads to the possibility that two players or groups of players will bid for the allegiance of a player by offering adjustments in the payoff schedule, or side payments.

All of this sounds very enticing and interesting, and I can just imagine some of you salivating and saying to yourselves, "Yeah, yeah, now he is getting to the good stuff." But I want to issue a caution. The appeal of Game Theory to social scientists, philosophers, and others, is that it offers a powerful analytical structure. That power is achieved, as I have labored to show you, by making a series of very precise, constraining simplifications and assumptions. As soon as you start relaxing those assumptions

and simplifications, you rapidly lose the power of the analytical framework. **You cannot have your cake and eat it too.** By the time you have loosened things up enough so that you can fit your own concerns and problems into the Game Theory conceptual framework, you will almost certainly have lost the rigor and power you were lusting after, and you are probably better off using your ordinary powers of analysis and reason. Otherwise, you are just tricking your argument out in a costume, in effect wearing the garb of a Jedi knight and carrying a toy light saber to impress your children.

3.2.3 Abrogating one of the Six Axioms

The six Axioms laid down by von Neuman conjointly permit us to represent a player's preferences by means of a cardinal utility function. There are various ways in which we might ease those axioms. One is to assume only an ordinal preference structure. As we have seen, that is sufficient for solving some two-person games, and it might be sufficient for usefully analyzing some multi-party games. We may need no more than the knowledge of the order in which individuals rank alternatives. All majority rule voting systems, for example, require only ordinal preference orders, a fact that is important when considering the so-called "paradox of majority rule."

The assumption of completeness is very powerful and potentially covertly biased in favor of one or another ideological position, a fact that I will try to show you when we come to talk about Nozick's work. In effect, the assumption of completeness serves the purpose of transforming all relationships into market exchanges, with results that are very consequential and, at least for some of us, baleful.

Transitivity is also a powerful assumption, and some authors, most notably Rawls, have chosen to deny it in certain argumentative contexts. Recall my brief discussion of Lexicographic orders. When Rawls says that the First Principle of Justice is "lexically prior" to the Difference Principle (Rawls, 1971), he is denying transitivity. He is also, as we shall see, making an extremely implausible claim. Whether he understood that is an

interesting question.

One of the trickiest thickets to negotiate is the relationship between money and utility. Because the Axioms we must posit in order to represent a player's preferences by a cardinal utility function are so daunting, those who like to invoke the impressive looking formalism of Game Theory almost always just give up and treat the money payoffs in a game [or a game like situation] as equivalent to the players' utilities. This is wrong, and some folks seem to know that it is wrong, but they almost never get further than just making some casual assumption of declining marginal utility for money. The issue of aversion to risk is usually ignored, or botched.

To give you one quick example of the tendency of writers to ignore the complexity of the six Axioms, here is the entry in the end-of-volume Glossary for "von Neuman-Morgenstern Expected Utility Theory," in *Game Theory and the Law* by Douglas G. Baird, Robert H. Gertner, and Randal C. Picker:

> Von Neuman and Morgenstern proved that, when individuals make choices under uncertainty in a way that meets a few plausible consistency conditions, one can always assign a utility function to outcomes so that the decisions people make are the ones they would make if they were maximizing expected utility. This theory justifies our assumption throughout the text that we can establish payoffs for all strategy combinations, even when they are mixed, and that individuals will choose a strategy based on whether it will lead to the highest expected payoff.

Now that you have sweated through my **informal** explanation of each of the six Axioms, I leave it to you whether they are correctly characterized as "a few plausible consistency conditions."

3.2.4 Relaxing the Assumption of Strictly Competitive Preferences

As I have already pointed out, there are a great many two-party situations [like two people negotiating over the price of a house] in which the parties do not have strictly opposed preference orders. This is manifestly true in nuclear deterrence strategy situations in which it is in the interest of both parties to avoid one outcome—namely mutually destructive all out war.

In addition to games that are partly competitive and partly cooperative, we can also consider totally cooperative games, sometimes called "coordination games." Here is one example. In his book, The Strategy of Conflict, Schelling cites a coordination game he invented to try out on his Harvard classes. He divided his class into pairs of students, and told them that without consultation, they were to try to coordinate on a time and place where they would meet. Each member of the pair was to write a time and place on a slip of paper, and then the two of them would read the slips together. "Winning" meant both students choosing the same time and place. An impressive proportion of the pairs, Schelling reported, won the game by coordinating on "Harvard Square at noon when classes let out." Obviously, their success in coordinating involved their bringing to the game all manner of information that would be considered extraneous in a competitive game, such as the fact that both players are Harvard students. Some time after reading this, I was chatting with a Harvard couple I knew, and I decided to try the game out on them. When I opened the first piece of paper, my heart sank. The young man had written, "4:30 p.m., The Coffee Connection." "Oh Lord," I thought, "he didn't understand the game at all." Then I looked at the young lady's piece of paper. It read, "4:30 p.m., The Coffee Connection." It seems that is where they met every day for coffee. Schelling wins again!

Not much in the way of theorems, but a great deal in the way of insight, can be gained from analyzing these situations, as Schelling has shown.

3.2.5 Mixed Strategies

The subject of mixed strategies has an interesting history. During the Second World War, the Allies struggled with the problem of defending the huge trans-Atlantic convoys of military supply ships going from the United States to England against then terrible depredations of the Nazi wolf packs of u-boats. The best defense was Allied airplanes capable of spotting u-boats from the air and bombing them, but the question was, What routes should the planes fly? if they planes, day after day, flew the same routes, the u-boats learned their patterns and maneuvered to avoid them. There was also the constant threat of espionage, of the secret anti-u-boat routes being stolen. The Allied planners finally figured out that a mixed strategy of routes determined by a lottery rather than by decision of the High Command held out the most promising hope of success.

Generally speaking, however, mixed strategies are a bit of arcana perfect for proving a powerful mathematical theorem but not much use in choosing a plan of action.

3.2.6 Calculation of Mathematical Expectation versus Maximization of Security Levels

We have already discussed at some length the limitations of maximization of expected utility as a criterion of rationality of decision making. von Neuman and Morgenstern reject it in favor of the much more conservative rule of maximizing one's security level. We have also seen that this rule of decision making does not allow for risk aversion [or a taste for risk], unless we totally change the set over which preferences are expressed, so that they become compound lotteries over even total future prospects rather than Outcomes in any ordinary sense. As we have also seen, maximization of expected utility rules out lexicographic preference orders, and when I come to talk about the application of this methodology to nuclear strategy and deterrence policy, I will argue that the assumption of non-lexicographic preference orders covertly constitutes an argument for a nuclear strategy favoring the Air Force or the Army rather than the Navy in the inside-the-Beltway budget battles.

3.2.7 Pre-Play Communication

Once we permit pre-play communication, all manner of fascinating possibilities open up. As we might expect, situations with pre-play communication and non-strictly opposed preference orders are among the richest fields for discussion and at the same time allow for the least in the way of rigorous argument or proof. In the hands of an author with a good imagination and a sense of humor, this can be lots of fun, but virtually everything that can be said about such situations can be said without calling them games and drawing imposing looking 2 x 2 payoff matrices. For example, as any hotshot deal maker in the business world knows, when you are engaged in a negotiation, it is sometimes very useful to make yourself deliberately unreachable as the clock ticks on toward the deadline for a deal. If a deal must be struck by noon on Tuesday, and if both parties want to reach agreement somewhere in the bargaining space defined by the largest amount of money the first party is willing to pay and the smallest amount the second party is willing to accept, it is tactically smart for the buyer to make a lowball offer within that space, and then be unavailable until noon Tuesday [somewhere without cell phone coverage, in the ICU of a hospital, on an airplane.] The seller must then accept the offer or lose the sale. Since by hypothesis the seller is willing, albeit reluctant, to sell at that price, she will accept rather than lose the sale. If the seller sees this coming, she can in turn give binding instructions to her agent to accept no offer unless there is the possibility of a counteroffer before the deadline. Then she can make herself unavailable. And so forth. This is the stuff of upscale yuppie prime time tv shows. It just sounds more impressive when you call it Game Theory.

3.2.8 Perfect Information

The general subject of perfect and imperfect information has been so much discussed in economics of late that I need not say anything here. Suffice it to note that formal Game Theory assumes perfect information of the payoff matrix, which embodies both the rules of the game and players' preference structures. Games do allow for imperfect information, of

course. Poker players do not know one another's cards, for example. But that is a different matter, built into the rules of the game.

Part III

Applications

Chapter 4

Applications

The time has come to put all of this formal stuff to use. In the second major part of this tutorial, I shall examine a number of attempts to apply the materials of Game Theory and Rational Choice Theory to substantive issues in political theory, economics, military strategy, and the law. My message will in the main be negative. I shall argue, again and again, that authors attempting to gain rigor or clarity or insight by the use of these methods actually misuse them, failing to understand them correctly or failing to understand the scope and nature of the simplifications and abstractions that are required before the materials of Game Theory and Rational Choice Theory can be properly applied. I have asked you to read two essays and a chapter of a book, all by me, and all available by clicking on the links provided in the blog post of June 2, 2010. In order to move things along and keep this tutorial to a manageable size, I am going to rely on you to do that reading, so that I can refer to it without summarizing it or repeating what I have said in those texts. My order of discussion will be as follows:

1. A discussion of the Prisoner's Dilemma

2. A discussion of the Free Rider Problem

3. An extended and very detailed analysis of the central thesis of John Rawls' *A Theory of justice*.

4. A brief discussion of certain arguments in Robert Nozick's *Anarchy, State, and Utopia* (Wolff, 1977b).

5. A discussion of some of the applications of Game Theory and Rational Choice Theory in *Game Theory and the Law* by Baird, Gertner, and Picker.

6. A discussion of the role played by Game Theory in the debates about military strategy and deterrence policy in the United States in the first twenty years following World War II. In connection with this portion of the discussion, I will make available the text of a book I wrote in 1962 but was never able to get published.

Assuming anyone is still with me after all of that, I will entertain suggestions of how we might usefully keep this tutorial going. Alternatively, I can go back to playing Spider Solitaire on my computer. :)

4.1 The Prisoner's Dilemma

The Prisoner's Dilemma is a little story told about a 2 x 2 matrix. For those who are unfamiliar with the story [assuming someone fitting that description is reading these words], here is the statement of the "dilemma" on Wikipedia:

> Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated the prisoners, visit each of them to offer the same deal. If one testifies for the prosecution against the other (defects) and the other remains silent (cooperates), the defector goes free and the silent accomplice receives the full 10-year sentence. If both remain silent, both prisoners are sentenced to only six months in jail for a minor charge. If each betrays the other, each receives a five-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?

The following matrix is taken to represent the situation.

	B1 cooperate	B2 defect
A1 cooperate	6 months, 6 months	10 years, Go free
A2 defect	Go free, 10 years	5 years, 5 years

Table 4.1: Payoff matrix for the Prisoner's Dilemma.

The problem supposedly posed by this little story is that when each player acts rationally, selecting a strategy solely by considerations of what we have called dominance [A2 dominates A1 as a strategy; B2 dominates B1 as a strategy], the result is an outcome that both players consider sub-optimal. The outcome of the strategy pair [A1,B1], namely six months for each, is preferred by both players to the outcome of the strategy pair [A2,B2], which results in each player serving five years, but the players fail to coordinate on this strategy pair even though both players are aware of the contents of the matrix and can see that they would be mutually better off if only they would cooperate.

For reasons that are beyond me, this fact about the matrix, and the little story associated with it, is considered by many people to reveal some deep structural flaw in the theory of rational decision making, akin to the so-called "paradox of democracy" in Collective Choice Theory. Military strategists, legal theorists, political philosophers, and economists profess to find Prisoner's Dilemma type situations throughout the universe, and some, like Jon Elster [as we shall see when we come to the Free Rider Problem] believe that it calls into question the very possibility of collective action (Wolff, 1990).

There is a good deal to be said about the Prisoner's Dilemma, from a formal point of view, so let us get to it. [Inasmuch as there are two prisoners, it ought to be called The Prisoners' Dilemma, but never mind.] The first problem is that everyone who discusses the subject confuses an outcome matrix with a payoff matrix. In the game being discussed here, there are two players, each of whom has two pure strategies. There are no chance elements or "moves by nature" [such as tosses of a coin, spins

of a wheel, or rolls of a pair of dice]. Let us use the notation O11 to denote the outcome that results when player A plays her strategy 1 and player B plays his strategy 1. O12 will mean the outcome when A plays her strategy 1 and B plays his strategy 2, and so forth. There are thus four possible outcomes: O11, O12, O21, O22.

In this case, O11 is "A serves six months and B serves six months." O12 is "A serves 10 years and B goes free," and so forth. Thus, the Outcome Matrix for the game looks like this:

	B1	B2
A1	A serves six months and B serves six months	A serves ten years and B goes free
A2	A goes free and B serves 10 years	A serves 5 years and B serves five years

Table 4.2: Outcome matrix for the Prisoner's Dilemma.

Notice that instead of putting a comma between A's sentence and B's sentence, I put the word "and." That is a fact of the most profound importance, believe it or not. The totality of both sentences, and anything else that results from the playing of those two strategies, is the outcome. Once the outcome matrix is defined by the rules of the game, each player defines an ordinal preference ranking of the four outcomes. The players are assumed to be rational—which in the context of Game Theory means two things: First, each has a complete, transitive preference order over the four outcomes; and Second, each makes choices on the basis of that ordering, always choosing the alternative ranked higher in the preference ordering over an alternative ranked lower.

Nothing in Rational Choice Theory dictates in which order the two players in our little game will rank the alternatives. A might hate B's guts so much that she is willing to do some time herself if it will put B in jail. Alternatively, she might love him so much that she will do anything to see him go free. A and B might be sister and brother, or they

might be co-religionists, or they might be sworn comrades in a struggle against tyranny. [They might even be fellow protesters arrested in an anti-apartheid demonstration at Harvard's Fogg Art Museum—see my other blog for a story about how that turned out.]

"But you are missing the whole point," someone might protest. "Game Theory allows us to analyze situations independently of all these considerations. That is its power." To which I reply, "No, you are missing the real point, which is that in order to apply the formal models of Game Theory, you must set aside virtually everything that might actually influence the outcome of a real world situation. How much insight into any legal, political, military, or economic situation can you hope to gain when you have set to one side everything that determines the outcome of such situations in real life?"

In practice, of course, everyone assumes that A ranks the outcomes as follows: $O21 > O11 > O22 > O12$. B is assumed to rank the outcomes $O12 > O11 > O22 > O21$. With those assumptions, since only ordinal preference is assumed in this game, the payoff matrix of the game can then be constructed, and here it is:

	B1	B2
A1	second, second	fourth, first
A2	first, fourth	third, third

[Notice, by the way, that this is not a game with strictly opposed preference orders, because both A and B prefer O11 to O22. With strictly opposed preference orders, you cannot get a Pareto sub-optimal outcome from a pair of dominant strategies—for extra credit, prove that. :)]

That payoff matrix contains the totality of the information relevant to a game theoretic analysis. Nothing else. But what about those jail terms? Those are part of the outcome matrix, not the payoff matrix. The payoff matrix gives the utility of each outcome to each player, and with an ordinal ranking, the only utility information we have is that a player ranks one

of the outcomes first, second, third, or fourth [or is indifferent between two or more of them, of course, but let us try to keep this simple.] But ten years versus going scot free, and all that? That is just part of the little story that is told to perk up the spirits of readers who are made nervous by mathematics. We all know that when you are introducing kindergarteners to geometry, it may help to color the triangles red and blue and put little happy faces on the circles and turn the squares into SpongeBob SquarePants. But eventually, the kids must learn that none of that has anything to do with the proofs of the theorems. The Pythagorean Theorem is just as valid for white triangles as for red ones.

To see how beguiled we can be by irrelevant stories, consider the following outcome matrix, derived from a variant of the story we have been dealing with:

	B1	B2
A1	A serves one day and B serves one day	A serves 40 years and a day and B goes free
A2	A goes free and B serves 40 years and a day	A serves 40 years and B serves 40 years

In this variant, if both criminals keep their mouths shut, they go free after only one night in jail. If they both rat, they spend forty years in jail. If one rats and the other doesn't, the squealer goes free today and the other serves 40 years and a day. Both criminals know this, of course, because the premise of the game is that this is Decision Under Uncertainty, meaning that they know the content of the outcome matrix and of the payoff matrix but not the choice made by the other player. The structure of the payoff matrix associated with this outcome matrix is supposed to be identical with that associated with the original story, namely: For A, $O21 > O11 > O22 > O12$, and for B, $O12 > O11 > O22 > O21$, because the premise of the little example is that each player rates the outcomes solely on the basis of the length of his or her sentence, regardless of how

long or short that is. It is therefore still the case that O11 is preferred by both players to O22, and it is still the case that IF each player's preference order is determined **solely** by a consideration of that player's sentencing possibilities [and that each player prefers less time in jail to more], and that each player chooses a strategy **solely** by attending to considerations of dominance, then the two of them will end up with a Pareto sub-optimal result. But how likely is all of that to occur in the real world? I suggest the answer is, not likely at all. For the upshot of the game to remain the same, we must assume two things, neither of which is even remotely plausible in any but the most bizarre circumstances: First, that each player is perfectly prepared to condemn his or her partner in crime to a sentence of 40 years and a day just to have a chance at reducing a one day sentence to zero; and second, that the two of them, faced with this extraordinary outcome matrix, cannot coordinate on the Pareto Preferred Outcome without the benefit of communication.

What would happen in the real world? I suggest something like this might happen: A examines the outcome matrix and says to herself: "Look, there is no difference to speak of between a 40 year sentence and a sentence of 40 years and a day. I am going to count on my partner to be sensible, and go for the one day sentence. The very worst that can happen is that I will have a day tacked onto the end of forty years, if I am still alive then, but I have a good shot here at getting off all but scot free."

Now, from a Game Theoretic point of view, this is not interesting at all. What is the point of introducing outcome matrices and payoff matrices and dominant strategies and Pareto sub-optimal outcomes if, when it gets right down to it, we are going to go into all the messy details of who the players are, what their relation to one another is, what history they have with one another, and all the rest of it? I thought Game Theory was going to enable me to analyze the situation without any of that stuff.

This is a point of such importance that I need to talk about it for a bit. A very long time ago, Aristotle and Pythagoras and some other smart Greeks [and also some really smart Egyptians, but I don't want to get into the whole Black Athena thing] discovered that in some situations, one can successfully abstract from the details of a problem and still carry out a

valid process of reasoning about it by attending only to certain formal or structural features of the situation. One can, for example, carry out long, complex chains of reasoning about shapes and sizes and spatial relationships without any reference to the materials in which these shapes and sizes and relationships are embedded. Now, this was not obvious on the face of it, when they made this historic discovery. You could not get very far reasoning about crops, after all, if you failed to take notice of which crop you were talking about, nor could you say much of interest about metalworking in abstraction from the particular metal in question. But if you know that all human beings are mortals, and you know that all Athenians are human beings, then you can draw the conclusion that All Athenians are mortal, just by attending to the formal syntactic structure of your two premises, namely that All A are B and All B are C, from which it follows, regardless of the details of the story you are telling, that All A are C.

Formal reasoning of this sort is beguiling, both because it is extremely powerful and because it can be engaged in by people who do not actually know much about the way the world works. There is also a lot of not very sublimated erotic and aggressive energy expressed here. Not for nothing do mathematicians speak about ramming an argument through. Oh well. That could lead us in rather hairy directions.

Once all of this has gained wide acceptance and has been brought to its present height of complexity and sophistication, everyone wants to get in on the act. I mean, who wants to talk about the psychological profiles of accused individuals enmeshed in the complexities of the criminal justice system when you can slap a 2 x 2 matrix on the page and carry out abstract calculations about dominant strategies? How cool is that? This is the reason why philosophers, who have long since learned that logicians have the highest status in their profession, put backwards E's on the page and talk about "for all x" rather than "everyone."

The little story called The Prisoner's Dilemma ignores just about every fact about a real Law and Order type situation that could possibly be relevant to thinking about it. Let us look at just a few of the things that are assumed away.

1. The situation is treated as a two person game. But there are obviously many more than two people involved. First of all, there are the cops who are putting the squeeze on the prisoners. In the real world, they are an important part of the situation, and real prisoners will try, quite rationally, to figure out whatever they can about the cops that will help them make their decision. Furthermore, in the American justice system, the prisoners will have lawyers. So at a bare minimum, this is a five-person game [one cop, two prisoners, two lawyers].

2. To force the story into a 2 x 2 matrix, one must suppose that each player has only two strategies. Recall what I said about how extraordinarily simple a game must be to offer only two strategies to each player. In the real world, there will be an arraignment, and there will be some jockeying over venue and date of trial and which judge is going to hear the case and whether to opt for a jury trial or go for a bench trial. Lots of moves, therefore lots of strategies, therefore no 2 x 2 matrix.

3. To make the story fit the matrix ["the punishment fit the crime"], we must abstract from every important fact about the two criminals, including sex, race, religion, personal relationship, past history with the criminal justice system, and so on and on, and then we must assume, against all plausibility, that each criminal will rank the outcomes purely on the basis of the length of the jail sentence to himself or herself.

Now, if we could, by doing all of this, draw conclusions whose validity is totally independent of all the details we have abstracted from, just as the validity of geometric calculation is independent of the color of the shapes whose area we are computing, then we would indeed have a very powerful tool for the analysis of economic, political, legal, and military problems. It would be a tool that could both help us to predict how people will act and also enable us to prescribe how rational individuals should act. But in fact, what remains when we have stripped away all the detail necessary

to reduce a complex situation to a 2 x 2 matrix is a structure that neither assists in prediction nor guides us in prescription.

If we focus simply on the formal structure of a two person game with two pure strategies for each player, it is obvious that there are 24 different orders in which each player can rank the four outcomes, setting to one side for the moment the possibility of indifference. How do I arrive at this number? Simple. A [or B] has four choices for the number one spot in the ranking. For each of these, there are three possibilities for the number two spot. There are then two ways of choosing among the remaining two outcomes for the number three spot, at which point the remaining outcome is ranked number four. $4x3x2x1 = 24$. Since A's rankings are logically independent of B's rankings, there are $24x24 = 576$ possible combinations of rankings by A and B of the outcomes of the four possible strategy pairs. The Prisoner's Dilemma is simply one of those 576, to which a story has been attached.

People enamored of this sort of thing have thought up little stories for some of the other possible pairs of rankings. [The following examples come from the pages of Baird, Gertner, and Picker, mentioned earlier]. For example, the following pair has had attached to it a story about The Battle of the Sexes [now fallen into disfavor for reasons of political correctness]:

A: $O21 > O12 > O22 > O11$
B: $O21 > O12 > O11 > O22$

Another pair of preference orders has a story about collective bargaining attached to it:

A: $O21 > O11 > O12 > O22$
B: $O12 > O11 > O21 > O22$

If we allow for indifference, then there are lots more possible pairs of preference orders. Here is one that has a story attached to it called The Stag Hunt:

A: $O11 > O21 = O22 > O12$
B: $O11 > O12 = O22 > O21$

I have no doubt that with sufficient time and imagination, one could think up many more stories to attach to yet other pairs of ordinal rankings of the four outcomes in a game with two pure strategies for each player. None of these little preference structures really models, in a useful way, relations between men and women, or collective bargaining, or stag hunts [since matching pennies really is a game, with all the simplifications and rules and such that characterize games, there is no reason at all why a Game Theoretic analysis should not be useful in understanding it, but one doesn't often encounter real world situations, even in Las Vegas casinos, where people are engaged in matching pennies.]

What is the upshot of this rather bilious discussion of The Prisoner's Dilemma? Put simply, it is this: The abstractions and simplifications required to transform a real situation of choice, deliberation, conflict, and cooperation into a two-person game suitable for Game Theoretic analysis fail to identify formal or structural features of the situation that are, at one and the same time, essential to the nature of the situation and independent of the facts or characteristics that have been set aside in the process of simplification. That, after all, is what does happen when we reduce an informal argument to a syllogism. Consequently, anything we can infer from the formal syllogistic structure of the argument must hold true for the full argument, once the content we have abstracted from is reintroduced.

Just to make sure this point is clear: Suppose I come upon a text in which the author tries to establish that some Republicans are honorable. She begins, we may suppose, by noting that all Republicans are Americans, and then offers evidence to support that claim the some Americans are honorable, whereupon he concludes that some Republicans are honorable. When we convert this to syllogistic form, it becomes: All A are B. Some B are C. Therefore, Some A are C. Thus separated from its content, the argument is quickly seen to be invalid [although, let us remember, that fact does not imply that the conclusion is false, only that it has not been established by the argument. Fair is fair.] The As that are B may not be

among the Bs that are C. [Venn diagrams, anyone?] In this case, the abstraction required to convert the informal argument into syllogistic form succeeds in identifying a formal structure of the original argument. Hence the formal analysis is valid.

But in the case of the Prisoner's Dilemma, essential elements of the original situation must be simplified away, removing aspects of the situation that are structurally essential to it. The result is not to lay bare the underlying formal structure of the original situation, but rather to substitute for the original situation another, simpler situation that can be exhibited in appropriate Game Theoretic form. The reasoning concerning this new situation is correct, but there is no reason to suppose that it applies as well to the original situation.

Conclusion: Be not beguiled by 2 x 2 matrices.

4.2 John Rawls' *A Theory of Justice*

We come now to what might plausibly be considered the real payoff for all the technical thrashing about we have been engaged in: an extended analysis of the core argument in John Rawls' famous book, *A Theory of Justice* (Rawls, 1971). Rawls' *hauptwerk* is widely considered the most important contribution to English language political theory of the past century, and is arguably the most influential work of philosophy written in the English language during that time. It is worth our while, therefore, to take the time to look at his central argument carefully and in detail. Thirty-three years ago I wrote a book-length examination of *A Theory of Justice*, called *Understanding Rawls*, published by Princeton University Press. Much of what I say here overlaps with what I said in that book (Wolff, 1977a), but my focus here is more narrowly on Rawls' attempt to apply Bargaining Theory to his subject. Those interested in a somewhat broader discussion are invited to hunt up my book and take a look. I am going to assume that everyone reading these words has some familiarity with Rawls' theory.

The core of Rawls' work is a simple and rather lovely idea. In the middle of the twentieth century, Anglo-American ethical theory was stuck

in what Kant, two centuries earlier, had called an antinomy. Utilitarians and intuitionists were locked in a death struggle, with each side more than capable of exposing the weaknesses of the other, but each unable to defend itself against the other's crushing arguments. Rawls had the idea that the conflict might be resolved by combining an old tradition of political philosophy—social contract theory—with a brand new field of mathematics and economics, Game Theory. Early in the development of the theory that eventually found its full-scale exposition in *A Theory of Justice*, Rawls claimed that he could prove a *theorem* in Bargaining Theory, and that the proof of that theorem would constitute a justification for the pair of principles which, he said, were or ought to be the foundation of a just society.

This was a very bold claim, and had Rawls been able to fulfill its promise, it would have been a monumental achievement. As we shall see, Rawls very early recognized that the original version of the theorem was unprovable, and indeed false. In response to this realization, he made sweeping changes to his theory, resulting in the distinctive form that the theory takes in *A Theory of Justice*, but unfortunately, the revised theory is not more defensible than the original. Rawls himself seems to have realized this fact, for while repeating the language of "theorem" and "proof," he very considerably backs away from the strong claims that he made in the earliest published version of his theory.

Before we begin the detailed examination of the argument, let me take just a moment to explain why I believe it is appropriate to bring the tools and insights of Game Theory to bear on *A Theory of Justice*. That is, after all, not the customary manner in which we engage with the arguments of Hobbes, Locke, Rousseau, Mill, Kant, or any of the other great figures of the Western tradition of democratic political theory. Quite simply, the reason is this: A great part of the plausibility of Rawls' theses derives from his claim that they can be grounded in a formal argument of Bargaining Theory. Absent that claim, the reader is left simply to contemplate Rawls' political theory and consider whether he or she likes it. The *argument* for the theory is, when all is said and done, the claim that the two principles would be chosen by rationally self-interested individuals situated in what

Rawls eventually came to call the Original Position. If that is simply not true, then it is hard to see what other justification Rawls has for his theory.

It is actually rather difficult to figure out exactly what Rawls' Two Principles *mean*, and the only way I can see to grapple with them is to take Rawls at his word that they are the solution to a bargaining game, and then see how we might so construe them. In this case, as we shall see, the formal machinery of Game Theory is quite helpful in guiding us to turn Rawls' non-technical language into something precise enough to be subjected to analysis.

It will be useful for our purposes to begin with the earliest statement of Rawls' theory, as it appeared in an article entitled "Justice as Fairness," published in 1962 in an important collective volume of essays called 1Philosophy, Politics, and Society, Second Series, edited by Peter Laslett and W. G. Runciman. Two passages from the essay will set things up for the first stage of my analysis.

"The conception of justice which I want to develop," Rawls writes, "may be stated in the form of two principles as follows: first, each person participating in a practice, or affected by it, has an equal right to the most extensive liberty compatible with a like liberty for all; and second, inequalities are arbitrary unless it is reasonable to expect that they will work out for everyone's advantage, and provided the positions and offices to which they attach, or from which they may be gained, are open to all."

These principles, Rawls says, would be agreed upon unanimously in a deliberation that he characterizes roughly in the way that "state of nature" political theorists describe the agreement on the Social Contract that constitutes a nation. Although he acknowledges that his remarks "are not offered as a rigorous proof" that persons engaged in this deliberation would agree on the two principles, such a proof requiring "a more elaborate and formal argument," nevertheless, he goes on to say:

> [T]he proposition I seek to establish is a necessary one, that is, it is intended as a theorem: namely, that when mutually self-interested and rational persons confront one another in typical circumstances of justice, and when they are required by a pro-

> cedure expressing the constraints of having a morality to jointly acknowledge principles by which their claims on the design of their common practice are to be judged, they will settle on these two principles as restrictions governing the assignment of rights and duties, and thereby accept them as limiting their rights against one another.

I think it is patently clear that Rawls in these passages is laying claim to the rigor and demonstrative power of Game Theory, at least in its somewhat looser form as Bargaining Theory. He seeks to show that his proposition is *a necessary one*; he asserts that it is intended as a *theorem*. Throughout the long evolution and transformation of his theory, Rawls never gave up this claim, for all that he also never came close to providing an argument for it. It is, I believe, the heart and soul of his entire enterprise. Without it, he has nothing but a rather affecting, albeit extremely murky, expression of his personal preferences in social organization.

What can Game Theory tell us about Rawls' claim? There are two questions that we must try to answer: What do his two principles mean? and Is his assertion a theorem that can be proved with necessity?

First, a problem: In the original statement of the principles, Rawls states his first principle thus: "each person participating in a practice, or affected by it, has an equal right to the most extensive liberty compatible with a like liberty for all. " In *A Theory of Justice*, however, the reference to practices is omitted, and instead we get "Each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others." [p. 60] Eventually, this is tweaked a bit, and becomes "Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all." [p. 302] The original formulation is thus intended to apply to practices, such as marriage, or capitalism, or the military, or the judicial system. The final formulation applies to nothing less than the total organization of a society. The shift, as we shall see, makes it forbiddingly difficult to figure out what the principle actually means.

The first principle, in all of its variants, uses the phrase "the most exten-

sive." That implies that one can rank alternative arrangements of a practice, or, alternative sets of fundamental or constitutional arrangements, in order of the degree of liberty that they embody or promise or make possible or guarantee. But as the term "liberty" is ordinarily used in the context of debates about political systems, it refers to a wide variety of institutional arrangements or guarantees that vary along multiple dimensions. The right to trial by a jury of one's peers, we may suppose, is a form of liberty. So is the right of all adults to vote periodically in elections to select the members of the government. Is a system of government with the first but not the second a more extensive or a less extensive liberty than a system of government with the second but not the first? One might reply, it does not matter, because a system of government with both is superior to either. Hence it would be Pareto preferred to either. But suppose there are liberties that in some of their forms are incompatible, in the sense that guaranteeing one interferes with guaranteeing the other.

For example [if I may be a trifle facetious], imagine a society consisting solely of authors and literary critics [these days, what with the internet, it does seem as though everyone is either an author or a critic, and sometimes both]. Now authors wish to be free of what they consider unfounded attacks on their writings, and critics wish to be free of what they consider unjustified limitations on their critiques. So the critics will prefer the American system of libel law, which gives very broad latitude to critics, and authors will prefer the British system, which favors those supposedly libeled [unless, of course, the authors wish, in their writings, to say nasty things about the critics, in which case the situation is more complex.] There is no way in this situation simultaneously to maximize the liberty of both groups, since each extension of the liberty of one will be experienced as a loss of liberty by the other. Consequently, as the authors and critics gather to bargain on the founding principles of liberty in their society, they will find it impossible to achieve unanimity on Rawls' first principle, because every attempt to spell out what it means will have them at loggerheads. This problem is not at all trivial, for all that my example may make it seem so. It is referred to in economics as the Indexing Problem, and it will crop again in a different guise in the final version of Rawls'

theory.

The second principle says that "inequalities are arbitrary unless it is reasonable to expect that they will work out for everyone's advantage." Remember that in the original version of Rawls' principles, it is inequalities in a practice that are being referenced—differential salaries, for example, or differences in the perks associated with one of the roles in the practice. The core idea is that economic inequalities may actually result in an increase in total social output, for example by attracting especially talented people to positions demanding highly skilled workers. Inequalities may also motivate young people to acquire time consuming and demanding skills whose deployment in more highly compensated positions will once again increase total output. Higher wages are required to attract the talented workers or to persuade them to spend the time and money acquiring the skills. If there is something left over from the increased output after the skilled workers have received sufficient additional compensation to motivate them, then the remainder—what we might call an "inequality surplus"—can be spread around to the rest of the society, making everyone better off than he or she would have been in a society that enforces equal compensation at the price of a universally lowered standard of living.

This is really the core idea in Rawls' entire theory of social justice. It is, we may note, the standard Sociological rationale for the extreme inequality of modern capitalist society. I like to think of it as the Brain Surgery argument—to wit, "If you have to have brain surgery, do you want to be operated on by someone who is paid no more than a burger flipper at Wendy's?" The idea roughly is this: If everyone were paid the same wage—say something above what is now Minimum Wage—no one would have any particular motivation to swap the job of burger flipper or ditch digger or garbage collector for the job of corporate executive or brain surgeon or professor, assuming that those jobs had been stripped of their various perks as well as of their higher salaries [corner offices, people who call you "sir" or "ma'am," and so forth.] But society needs talented corporate executives and well-trained brain surgeons and professors. Otherwise the Gross Domestic Product will be sub-optimal and everyone will suffer. So efficiency demands that we slap big salaries on those jobs to

motivate some of the more promising burger flippers and garbage collectors to trade in their jobs for brain surgery. If we manage things skillfully, we will pay them just enough to drag them away from their spatulas and garbage trucks and into the executive suites and hospitals. Those salary increases will come out of the GDP, of course, but there will be enough left over to raise the pay of all the remaining burger flippers and garbage collectors. So, assuming no one is envious, and resents the higher salaries of the brain surgeons even though the productivity of the brain surgeons is raising the wages of the burger flippers, everyone will be in favor of this inequality. [This is the reason for the non-envy clause in Rawls' theory, in case you ever wondered.]

I have spelled this out at length because thus explicated it is so wildly implausible. As a description of what motivates people in a modern capitalist society to pursue one career path rather than another it is so tone-deaf sociologically and psychologically as to sound like a Jon Stewart send-up. But that is not the focus of these remarks, so let us leave that to one side.

One of the purposes of the Difference Principle, so called, is supposed to be to allow us to adjudicate complaints against a scheme of unequal compensation by showing that the inequality works to everyone's advantage. But "to everyone's advantage" is, grammatically speaking, a comparative rather than a superlative. To everyone's advantage compared to what? This is a good deal harder to answer than it might seem at first glance.

Presumably, the answer is that the scheme works to everyone's advantage compared to the same social system with equal compensation, if indeed one can even imagine the same social system without the inequalities. [Would a corporation be the same institution if everyone made the same wage? Would a hospital be?] But there are almost certainly a number of alternative schemes of unequal wages, each of which generates an Inequality Surplus adequate to make everyone better off, **and each of which makes some positions better off and some positions worse off than those positions would be under a different inequality scheme generating a surplus**. If that is so, considerations of Pareto Comparability do

not permit us to say which scheme is preferable to which, even though each of them is preferable to the society without inequality.

Assuming [what is in fact false] that all of these problems with the Two Principles can be solved, we are left with the central question: Would a group of rationally self-interested individuals faced with the circumstances of justice necessarily coordinate on the Two Principles? [By the way, for those who are not up to speed on all of this, the "circumstances of justice" are these: First, that the members of the society have something to gain from cooperation, if they can only agree; and Second, that their pre-agreement assets and powers are sufficiently equal to motivate them to seek common agreement. This is all standard Social Contract stuff, straight out of Hume, Hobbes, etc.] In short, we are faced with a proposed theorem in Bargaining Theory.

As Rawls conceives the Bargaining Game, this is a multi-party game with full communication among players who are assumed to have cardinal utility functions invariant under affine transformations. Rawls never tells us how the game is played, nor does he even seem to think that he needs to do so. That is one of the odd things about his invocation of "theorems" in Game Theory. We are left to try to imagine for ourselves how the game would actually be played. I think we are meant to imagine something like this: The players sit in a circle in such a way that what each says is heard, and is known to be heard, by all. One player starts, and proposes a rule. [Say, the old Bill Cosby rule from early Sesame Street—"All for one and everything for myself."] The next player either accepts the first player's rule, which of course she won't, or proposes a new rule. They go around the circle again and again, proposing foundational rules, until they succeed in making one complete circle during which everyone agrees to the same rule. That is then the solution to the game. Rawls says that after a bit "they will settle on [his] two principles." Is he right?

Alas, no. There are two problems, one procedural, the other substantive. The procedural problem is that the bargaining game has no termination rule. There is no reason for the first player ["Everything for myself"] ever to stop proposing that rule. There is presumably some very, very small, but not zero, probability that sooner or later [probably later] the

rest of the players will get tired or drop the ball and agree. Since there are no costs in the game associated with continuing to play it, none of the players has any incentive to "be reasonable."

You can fix this glitch, of course, by imposing a time limit on the game. But that gives an asymmetrical advantage to the last player whose turn leaves just enough time to go around the circle once. When that moment is reached, the lucky player can propose a rule that makes everyone better off than having no rule at all, but advantages that player [such as "People with naturally curly hair get first dibs on all the nifty jobs," said by someone who has naturally curly hair]. Now, this is silly, right? But when you claim to be proving a theorem that is necessary, that is the sort of thing you have to take into account. This is an example of what I mean by wrapping yourself in the impressive language of Game Theory to make what you are doing sound impressive, while not actually engaging in Game Theoretic arguments. As Rawls says in "Justice as Fairness," "there remain certain details to be filled in, and various alternatives to be ruled out." Indeed.

The second problem is substantive. The two principles proposed by Rawls would not win unanimous agreement from the players. The problem is this: While the players are faced by the circumstances of justice, and hence are roughly comparable in their powers and endowments, there are nevertheless significant differences among them in natural talents and abilities. Some of them will fare much better than others in a society in which "the positions . . . are open to all." What is more, despite these differences in native intelligence or ability, all of them know these facts. Now, if you are one of the most talented members of society, you are going to be in favor of a structure of inequality in which you know quite well that you will be one of those ending up in the favored positions. But if you are not one of the talented, you will conduct an expected utility calculation and come to the conclusion that you might be better off with a system in which the better paid jobs are distributed by lot to anyone who meets certain minimum requirements [for example, a college degree, even with a low GPA]. The imposition of minimum requirements will suffice to generate some Inequality Surplus, but the allocation of the favored jobs by lot will work to the advantage of the less talented members of the society, who will thus

have a shot at the higher paid jobs, something they would never have if those jobs were allocated strictly on the basis of a fair competition. [As Senator Roman Hruska of Nebraska said in 1970, defending the Supreme Court nomination of G. Harrold Carswell from the charge of mediocrity, "There are a lot of mediocre judges and people and lawyers. They are entitled to a little representation, aren't they?"]

So Rawls' "theorem" is no theorem at all. I pointed this out in an article I published in *The Journal of Philosophy*, and Jack's face fell when I told him about it at an annual APA meeting. But I went on to say that his subsequent essay, "Distributive Justice," changed his theory in a way that met all of my objections, and his face brightened. "Oh, that's all right then," he said before wandering off.

The major changes to the bargaining game introduced in "Distributive Justice" and carried over into *A Theory of Justice* are rather dramatic. They are four in number:

1. Rawls introduces the famous Veil of Ignorance, which is a brilliant literary device designed to capture what we mean when we say that judges must be "disinterested"—i.e., that they must make decisions without taking into consideration their own interests or situation. This takes care of the problem that some participants in the Bargaining Game are, and know they are, among the more talented members of society, while others are, and know they are, among the less talented. But this fix comes with its own problems.

   Stripped of any knowledge of who they are, the players in the Bargaining Game now have no reason at all to bargain, or indeed to do anything else. This will be clear if we keep in mind the analogy to judges. Judges, in their judicial capacity, are not supposed to have an agenda [think Confirmation Hearings for Supreme Court nominees and all the blather about "activist justices" who "make law from the bench."] So if the individuals under the veil of ignorance know nothing at all about who they are, they can have no ends, no purposes, save perhaps the goal to render decisions fairly when they are presented with cases. Hence these individuals have no reason to

bargain about anything at all. The self-interest Rawls has equipped them with is vacuous. Rawls is thus forced to endow the individuals in the Original Position with some sorts of purposes, specific enough to make them care what they get from the agreement being hammered out, but not so specific as to recreate the problems that invalidated the first version of his theory.

2. Rawls' solution is to state that the individuals in the Original Position know that they have Life Plans. And since it wouldn't do to allow them to have the life plan of a religious hermit, for example [because then they would not care how much stuff they got out of the bargain, or even whether they had civil rights and protections], Rawls stipulates that the Life Plans posited for the individuals in the Original Position require for their accomplishment certain rights and abilities, powers and endowments [i.e., some stuff]. Bu this creates a new raft of problems, of an especially intractable sort, because the sorts of rights and abilities, powers and endowments required for the fulfillment of one Life Plan may be quite different from those required for the fulfillment of a different Life Plan. If your Life Plan is to become a Professor of Philosophy, then access to higher education for all those sufficiently talented is going to loom large in your budget of things you are bargaining for. But if your Life Plan is to become a champion surfer, higher education may drop way down in your list of desiderata.

   [Notice, by the way, that built into Rawls' conception of Life Plans is a particular substantive historical, and cultural, and social conception of individual personality and the good life. I do not have time to go into this in the detail that it deserves, so I will simply refer you to the classic discussion in Karl Mannheim's *Ideology and Utopia*, Chapter IV, of the chiliastic, liberal-humanitarian, conservative, and socialist-communist forms of the Utopian Mentality and the orientation to time itself. Rawls, to put it in shorthand, assumes that everyone in the Original Position has a liberal-humanitarian orientation toward time, which is simply part of the larger fact that his entire theory

is an ideological rationalization of capitalist social democracy. But I digress.]

To get over the problem posed by the diversity of possible Life Plans, so that he can get back to his theorem, Rawls now makes another assumption, and this one is, technically speaking, a whopper.

3. Since they do not know which particular Life Plans they have, Rawls asserts [he never offers anything remotely like an argument for it] that one can create an index of the heterogeneous basket of basic stuff that anyone would need to pursue whatever Life Plan he or she might turn out to have, something Rawls calls an Index of Primary Goods He then simply assumes that everyone in the Original Position has positive, albeit declining, marginal utility for the Index of Primary Goods. He assumes this, not because it is plausible [he never offers any arguments for any of this], but because unless he assumes it he won't have a hope of invoking the tools and techniques of modern economic theory with which he thinks he can prove his "theorem."

   The problem of constructing an index of a heterogeneous assortment of rights and abilities, powers, and endowments is completely insoluble, as any honest economist will tell you. Indices like the Consumer Price Index or the Dow Jones Index are hopelessly flawed, and there is no way to fix them. Some of you may be quite familiar with this problem, others may have never heard of it before. I guess maybe I ought to say a word or two about the subject, even though it will slow us down some. The Consumer Price Index is constructed by putting together a list ["a market basket"] of consumer goods in specific proportions that is supposed to reflect the way most Americans spend their household income in a specified time, say a month: So much housing, so many pounds of potatoes, so many lamb chops, so many visits to the doctor, etc. Samples are then taken, and averaged out, of what these items are selling for in various stores, doctors' offices, etc, around the country. The same market basket of goods and services is then priced a month later. If

the total cost of the market basket has increased by 1%, then it is said that there has been a 1% increase in the CPI, or that there has been a month to month inflation rate of 1%.

There are a gazillion problems with this index, all of which are totally unfixable. For example, suppose you have a thirty year fixed rate mortgage on your house. In that case, if housing costs rise dramatically [as they did in the 70's and 80's, for example], the Consumer Price Index, reflecting that rise, may shoot up dramatically. But you won't actually experience that rise, because a major component of *your* market basket of goods, namely housing, is fixed. The same sort of problem arises with regard to every element in the "market basket." The dramatic rise in health care costs will not affect you so long as you are healthy, but will affect you if you have a special needs child. if you are a vegetarian, a steady decline in the price of meat won't have any effect on your pocketbook. When I was a young man living in Cambridge, Massachusetts, lobster was cheap at the supermarket and steak was expensive. Now, when I shop for dinner, steak is cheaper per pound than all but the least expensive fish. Once again, Rawls unconsciously [I think] builds intro his supposedly universal theory the assumptions, presuppositions, and tastes of a particular social and economic class, which happens to be his.

4. The last major change is a revision in the statement of the so-called "Difference Principle." Instead of inequalities working to everyone's advantage, they must now work to the benefit of the least advantaged members of society. This is a very significant change However, Rawls introduces it *not* by saying that he has changed his mind, or has been compelled by the logic of the bargaining game to alter the theorem, *but rather* by saying that this is a more reasonable "interpretation" of the original form of the principle. I confess that I find this rather weird and creepy. Rawls treats his own two principles, *which he made up*, as though they had been inscribed on tablets by The Lord God Himself, and thus required us to interpret them rather

> than change them. This is one of the strangest features of Rawls' entire discussion.

Why the change? Pretty clearly, it is required by the special conditions imposed by the Veil of Ignorance. Since the parties in the Original Position now do not know who they are, they are pretty well forced *either* to carry out expected utility calculations over the entire range of positions in the society, some one of which each of them will actually turn out to occupy when they leave the Hall of Justice and regain their knowledge of who they are, *or else* to adopt the conservative assumption that they occupy the lowest position, and bargain to improve its allocation of Primary Goods. As we shall see, this is Rawls' version of the conservative rule proposed by von Neuman, to maximize one's security level.

All this goes by so quickly in Rawls' exposition of the mature form of his theory that unless you are paying real close attention, you may not notice how wildly implausible, or even downright impossible, it all is. This is the form of the theory that everyone is familiar with, but people usually do not have any coherent idea why Rawls has made all of these very powerful stipulations. The reason, as I have indicated, is that each element of the final theory is designed to meet an objection to an earlier form of the theory.

So where does all of this revision leave us? Well, first of all, something odd has happened along the way, as Rawls has altered his description of the choice situation to meet and overcome the difficulties with the first formulation. The original idea was that the parties to be governed by the agreed upon foundational rules would confront one another and bargain. The parties were assumed to be rationally self-interested, but with differing interests and desires. Rawls' central idea was that if to this premise of rational self-interest we added only one additional premise—the willingness of the parties to abide by a set of rules arising from the bargain, the willingness to take that one step beyond self-interest to something resembling what is involved in having a morality—then we could prove that the one and only set of rules on which they would self-interestedly settle would be his "two principles."

But if we think about it for a moment, we will realize that after the revisions, Rawls no longer has a Bargaining Game that looks anything like this. Since the players have been stripped of any individuating features that might distinguish them from one another [such as differences in tastes or talents, or indeed even differences in which stage of human history they happen to be located in], there are no rational grounds on which any two of them could reason differently from one another about the choice of the basic structure and rules of the society in which they will find themselves when they emerge from the Veil of Ignorance. In short, what began as a problem in Bargaining Theory has morphed into a problem in the Theory of Rational Choice. [This is one of the reasons why Rawls tended to move toward what he himself called the Kantian Interpretation" of his theory. But that really does get us too far afield.]

Before addressing the central question, viz., are these two principles thus revised, the solution to the Bargaining Game, thus altered, there is one subsidiary matter I should like to take up. Rawls says that although the parties in the Original Position under the Veil of Ignorance have temporarily forgotten who they are, what their specific desires are, and where they are located in history and in the structure of their society, they do retain a knowledge of the "general facts about human society." Each of these individuals, Rawls elaborates, understands "political affairs and the principles of economic theory," as well as "the basis of social organization and the laws of human psychology." In my book, *Understanding Rawls*, I argued that this is an epistemologically impossible state of affairs. There is not time or room here to repeat what I have said there [another shameless plug for my book :)], but I think it is worth indicating the line of argument that I develop there.

The first thing to be clear about is that under the Veil of Ignorance, the individuals in the Original Position do not even know in which stage of human history they are located. This fact leads Rawls into an extremely interesting discussion about the appropriate rate of savings that should be chosen as part of the basic socio-economic structure being negotiated. Debates about the social rate of savings are familiar to economists, but have been virtually absent from the political philosophy literature. It is greatly

to Rawls' credit that he recognized this and introduced the subject into his theory. For those who are unacquainted with the discussion, the central issue is this: The capital required for future economic activity [seed corn, machinery, Research and Development, and their monetary equivalents] must be obtained from current production by somehow imposing limits on consumption – eat all the corn this season, and there is no seed for next season's crop. Simple prudence dictates that people this year save for next year. But what shall we say about the responsibility of people in this generation to save for generations as yet unborn? A high, self-denying rate of social savings, such as that now being enforced by the Chinese government, will make possible an explosion of production in future generations, to the manifest benefit of those who are then alive. But that future production will come at the expense of this generation, which will have to deny itself some measure of present consumption.

From the perspective of Rawls' theory, the question becomes: Under the Veil of Ignorance, what rate of savings will rationally self-interested individuals choose to impose upon themselves once they emerge from the Veil and discover which generation of their society's evolution they are actually located in? I encourage readers interested in this subject to take a look at Rawls' discussion.

But getting back to the epistemological issue, the individuals in the Original Position are presumed to know the general facts of nature, society, economy, and human psychology, and even to know the broad outlines of the historical evolution of societies, but not to know where in that evolutionary process they are themselves located. Rawls clearly thinks it is possible for someone to be in this particular epistemological position. I think it is not. Why not?

First of all, the individuals in the Original Position are blocked from accessing certain individuating facts about themselves, but they have not lost their powers of reason. To put the point simplistically, if the Veil has enabled them to retain the knowledge that All men are mammals and the knowledge that All mammals are animals, then their unimpaired powers of reason will allow them to conclude that All men are animals. Somewhat more to the point, if they know the standard theorems concerning

the relation of supply to demand in the determination of price in a capitalist economy based on the production of commodities for sale in the marketplace, then they will be able to infer that their society has undergone the transition from Feudal to Capitalist social relations of production, *because until such a transition has taken place, individuals do not even possess the concepts that are employed in the formulation of those economic laws*. What is more, if, as I believe, capitalist social relations of production systematically mystify the underlying structure of exploitation on which capitalist profit rests, so that people mistakenly but inevitably perceive those relations as the expression of eternal and immutable economic laws, then only someone enmeshed in a capitalist society and economy will make the mistake of thinking that there are "laws of supply and demand."

Now, maybe I am right about that, and maybe I am wrong. But by building these assumptions into the structure of the bargaining game from which he hopes to extract **the** principles of justice, Rawls has begged all of the questions that might be raised by someone like me ["begged" in the proper use of that term—i.e., assumed what is to be proved]. This is one more example of my general claim that the misuse of formal methods allows authors to present their ideologically laden assumptions as value-neutral elements of a formal analysis.

Let us now return to the central question: Would the individuals situated under the Veil of Ignorance in the Original Position coordinate on Rawls' Two Principles of Justice as revised in *A Theory of Justice*? This question is much more difficult to answer now than it was with regard to the first form of the theory. Even to make the question determinate enough to grapple with it we must make a considerable number of assumptions and specifications with regard to matters that Rawls either does not discuss or else leaves up in the air.

At this point, in order to make this manageable, I must ask you to consult the chapter from my book, a link to which was posted earlier on this blog. I will discuss the problems in general terms, and leave it to each of you to read my detailed analysis in that chapter.

The first thing an individual in the Original Position must do when confronted with a choice of basic organizational rules for society is to

decide how well or badly off she is, or was, before entering the Hall of Justice. [I shall simply stipulate that our representative person is female, but of the course the person does not know this, or indeed anything else of a particular nature about him/her self.] Since Rawls says that she is rationally self-interested, and is prepared to enter into the bargaining game because she believes that a satisfactory outcome will be to her advantage, she clearly needs to know what her baseline situation is. Otherwise, she cannot make a judgment as to whether a proposed rule will make her better off. Remember: she not only does not know who in particular she is or where in her society she is situated. She also does not know what stage of history she is located in.

Faced with the necessity of stipulating a pre-bargain baseline [defined, we may suppose, simply by some specified amount of Primary Goods—this whole thing just gets hopelessly complicated if we try to flesh out her situation in any more realistic manner], she really has only three options. For each possible stage of history in which she might be located, she can either adopt the premise that she is the worst off representative person in that society; or she can adopt the premise that she is the best off representative individual; or she can carry out an expected utility calculation, assigning some level of Primary Goods and some probability to every representative position in the society, and then multiplying the two and summing the results, In this third case, she will say to herself something like this: "There are seven representative positions in the society; fifteen percent of the people are in the first, ten percent in the second, etc. The first position has so and so much of the Primary Goods assigned to it, the second such and such amount, and so forth; with no more information than that I am one of the people in the society, I conclude that I have a fifteen percent chance of being in the first position, a ten percent chance in the second position, and so forth. Assuming that I know what my cardinal utility function is for Primary Goods, I can now carry out my expected utility calculation."

Sigh. I told you this was going to be messy. I am pretty sure, from correspondence I had with Jack, that he is aware of a good deal of this, but I do not think he ever fully appreciated how deeply it undercut his

central claim that he was advancing a *theorem*. At this point, Rawls says that a rational person, recognizing how important the choice is that she is about to make, will adopt an extremely *conservative* way of evaluating alternatives. What does this mean?

Well, the first thing it means is assuming that outside the Hall of Justice, in the real world, she is one of the persons occupying the least advantaged representative position in society. Why is this conservative? Because if she assumes that she is in fact well off in the real world, she will be correspondingly less willing to make a deal, and this threatens to leave her utterly disadvantaged should the optimistic assumption about herself prove false. She must protect herself against the chance that she is one of society's poor, and the best way to do this is to agree to inequalities of any sort only if they work to the advantage of those least well off.

But reasoning in this fashion, she might be tempted to carry out some sort of expected utility calculation and opt for a set of principles that maximizes the average utility that each representative person will enjoy. To be sure, that can be risky, since a higher average overall might be compatible with a lower utility to the least well off. In an expected utility calculation, that risk might be compensated for by a chance at a very much higher payoff to the better off representative positions.

Rawls now argues that the rational individual under the Veil of Ignorance will reject expected utility calculations and instead opt for the extremely conservative, and also extremely controversial, "maximin" rule proposed by von Neuman. On page 163 of my book [see the chapter to which I have linked], I quote Rawls' reasons for adopting this rule. Here is what he says: "There are three chief features of situations that give plausibility to this unusual rule... The situation is one in which a knowledge of likelihoods is impossible or at best extremely insecure... The person choosing has a conception of the good such that he cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact, be sure of by following the maximin rule. It is not worthwhile for him to take a chance for the sake of a further advantage, especially when it may turn out that he loses much that is important to him... The rejected alternatives have outcomes that one can hardly accept. The situation involves

grave risks." [All four passages from Rawls, p. 154]

In my book, I have given a formal analysis of these claims, complete with nifty diagrams, but I want here to step back and try to get a sense of what Rawls is really talking about. Remember, first of all, that Rawls is not talking about the quantity of Primary Goods that the various principles of justice offer as possibilities, but rather about the utility that the *utility* function of the individual under the Veil of Ignorance associates with these various amounts of Primary Goods. The distinction is essential for understanding what Rawls is saying.

Concretely, Rawls is claiming that the rational individual under the Veil of Ignorance will say to herself: "If I opt for a system of social organization that holds out the possibility of vast wealth for a few, but that fails to protect those at the bottom from absolute penury, I am risking ending up in a disastrous situation, one that "involves grave risks." But all I stand to gain is the chance at one of the top spots, even though I "care very little, if anything, for what [I] might gain above the minimum stipend that [I] can, in fact, be sure of by following the maximin rule."

Fair warning: I am now going to say something that is mean-spirited and snarky, but I really do not know how else to get at what is going on in this argument. I apologize if I offend anyone. Here goes:

What sort of person says to himself or herself what the individual in the Original Position, according to Rawls, says? Not just a rational person. There is nothing formally irrational about being willing to risk utter penury for a chance at fabulous wealth. That is just a matter of having a utility function of a particular shape [one that is, over a certain range, monotonically increasing rather than decreasing.] Would Gordon Gekko think this way? [If there is anyone who does not recognize the name, Gordon Gekko is the main character of the 1987 film, *Wall Street*, starring Michael Douglas. If you haven't seen it, by all means get it from NetFlix.] Of course not. But Gordon Gekko is not formally irrational. He just places a very high value on vast wealth and has a very high tolerance for risk. What about Picasso? I think not. If you offered Picasso a chance at artistic immortality, with penury and misery as the alternative if he turned

out not to have real talent, I think he would have grabbed the chance with both hands. In fact, of course, he did.

No, the sort of person who would reason as Rawls thinks the individual in the Original Position would is a tenured professor—someone who has a comfortable albeit modest lifestyle that is absolutely assured against any risks, someone who has perhaps turned down other careers offering much larger rewards but also "involving grave risks." In short, the sort of person who would reason as Rawls thinks the individual in the Original Position would is . . . John Rawls.

Strip away all the talk about theorems, all the lovely filigree of philosophical elaboration, all the Reflective Equilibrium and Strains of Commitment and allusions to Game Theory, and you have a simple *apologia pro vita sua*.

If the Representative Individual in the Original Position is an academic at a good American university or college that offers life tenure and a comfortable middle class life, then I think it is quite likely that he or she would opt for Rawls' two principles. They guarantee a continuation of that pleasant life style, combined with a virtuous but really cost free concern for the poor downtrodden denizens of the Inner City [the least well off representative individuals].

Now, that is just about as mean-spirited as I have ever been in print [though not, I am afraid, in person], but what else can one conclude if one takes Rawls' theory seriously and tries to think through what it really means?

The time has come to step back from the details of Rawls' discussion and try to get some perspective on what is, when all is said and done, the most important contribution to political philosophy of the past hundred years and more. I observed at the beginning of these remarks that Rawls offered his very new theory at a time when Anglo-American Ethical Theory was mired in an antinomy—a several decades long face off between Intuitionism and Utilitarianism. Rawls invited us to get past that stalled historical moment by making use of ideas drawn from Game Theory [and also from neo-classical economics, but that is another matter.] If he had simply offered his Two Principles as an alternative to, or perhaps

more accurately as a fusion of the best parts of, Intuitionism and Utilitarianism, there is no question that his proposal would have commanded considerable attention. The elegance of his discussion of Utilitarianism and the interesting and suggestive detail of the fully elaborated version of his proposal would, I am sure, have generated a lively discussion among philosophers, political theorists, and others.

But what made Rawls' theory stand out as deserving of what constitutional lawyers call heightened scrutiny was his claim to be able to establish his two principles as the solution of a bargaining game. Now, even if this thesis could be sustained, it would still be open to readers to reject Rawls' claim that the solution of such a game ought to be considered *the principles of social justice*. But a genuine proof of Rawls' theorem would have vaulted his theory to an entirely unique status in ethical and political theory. Such a theorem would have taken its place beside Kenneth Arrow's General Possibility Theorem as a major result of formal analysis. [I remain convinced, in the absence of any textual or anecdotal evidence whatsoever, that this is exactly what Rawls dreamed of accomplishing.] This is why, both in my book and in these blog posts, I have focused almost exclusively on the logical status of the theorem that Rawls adumbrates in "Justice as Fairness," and continues to allude to as a theorem, albeit in a hedged manner, in "Distributive Justice" and *A Theory of Justice*.

I think I have demonstrated that the theorem is not valid, either in its original or in its revised form, or, more precisely, that it can only be made plausible by so many *ad hoc* adjustments, presuppositions, and qualifications that it loses its grip on our attention. I also think it is clear that the theory, as Rawls sets it forth in his book, covertly valorizes, without adequate argument, one particular substantive vision of the good society—a vision some components of which I share, but for which Rawls fails to offer an argument.

Well, this is twenty-four pages about Rawls, which is enough, I think, for this blog. I will turn my attention next to the single most important formal result in the application of formal methods to political philosophy: The General Possibility Theorem of Kenneth Arrow. My tone will change dramatically, as you will discover. No sniping or snarking, no *ad hominem*

arguments. Arrow's result, like von Neuman's Fundamental Theorem, is a genuine triumph, and I shall do my best in expounding it to make its logical structure clear.

4.3 Collective Choice Theory

Collective Choice Theory is the theory of how one selects a rule to go from a set of individual preference orders over alternatives available to a society of those individuals to a collective or social preference order over those same alternatives. [Or, as they say in the trade, how to "map a set of individual preference orders onto a social preference order."] There is a long history of debates about how to make social or collective decisions, going back at least two and a half millennia in the West. The simplest answer is to identify one person in the society and stipulate that his or her preference order will **be** the social preference order. *L'etat, c'est moi*, as Louis XIV is reputed to have said. A variant of this solution is the ancient Athenian practice of rotating political positions. One can also choose a person by lot whose preferences will thereupon become the social preference. A quite different method is that used by the old Polish parliament, which consisted of all the aristocrats in the country [there were quite a few, the entry conditions for being considered an aristocrat being low]. Since each of them thought of himself as answerable only to God, they imposed a condition of unanimity on themselves. If as few as one Polish aristocrat objected to a statute, it did not become law.

These rules for mapping individual preference orders onto a social preference order, unattractive as they may be on other grounds, all have one very attractive feature in common: They guarantee that if all of the individual preference orders are **ordinal orderings**, which is to say if each of them is complete, reflexive, and transitive [you see, I told you we would use that stuff], then the social preference order will also be an **ordinal ordering**, and that is something you really, really want. You want it to be complete, so that it will tell you in each case how to choose. And you want it to be transitive, so that you do not get into a situation where your **Collective Choice Rule** tells the society to choose a over b, b over c, and c

over a.

To sum it all up in a phrase, the aim of Collective Choice Theory is to find a way of mapping **minimally rational individual preferences onto a minimally rational social preference**.

For the past several hundred years, everybody's favorite candidate for a Collective Choice Rule has been **majority rule**. This is a rule that says that the social preference between any two alternatives is to be decided by a vote of all those empowered to decide, with the alternative gaining a majority of the votes being preferred over the alternative gaining a minority of the votes. Should two alternatives, in a pairwise comparison, gain exactly the same number of votes, then the society is to be **indifferent** between the two.

Enter the Marquis de Condorcet, who published an essay in 1785 called [in English] *Essay on the Application of Analysis to the Probability of Majority Decisions*. In this essay, Condorcet presented an example of a situation in which a group of voters, each of whom has perfectly rational preferences over a set of alternatives, will, by the application of majority rule, arrive at an inconsistent group or social preference. This is, to put it as mildly as I can, a tad embarrassing. Indeed, it calls into question the legitimacy of majority rule, which lies at the heart of every variant of democratic theory that had been put forward at that time, or indeed has been put forward since.

Let us take a moment to set out the example and examine it. In its simplest form, it involves three voters, whom we shall call X, Y, and Z, and three alternatives, which we shall call a, b, and c. We may suppose that a, b, and c are three different tax plans, say. Let us now assume that the three voters have the following preferences over the set of alternatives $S = (a, b, c)$.

X prefers a to b and b to c. Since X is minimally rational, he also prefers a to c.

Y prefers b to c and c to a. Since she is also minimally rational, she prefers b to a.

Z prefers c to a and a to b. As rational as X and Y, she naturally prefers c to b.

Now they take a series of pairwise votes to determine the collective or social preference order among the three alternatives. When they vote for *a* or *b*, X and Z vote for *a*, Y votes for *b*. Alternative *a* wins. When they vote for *b* or *c*, X and Y vote for *b*, Z votes for *c*, alternative *b* wins. Now, if the social ordering is to be **transitive**, then the society must prefer *a* to *c*. What happens when X, Y, and Z choose between *a* and *c*? X prefers *a* to *c*. But Y and Z both prefer *c* to *a*. So the society must, by majority rule, prefer *c* to *a*. Whoops. The society's preference order violates transitivity.

And that is the whole story. The selection of a social or collective preference order by majority rule cannot guarantee the transitivity of the social preference order, and therefore does not even meet the most minimal test of rationality. There are, of course, lots and lots of sets of individual preference orders that generate a consistent social preference order when Majority Rule is applied to them. The problem is that here is at least one, and actually many more, that are turned by Majority Rule into an inconsistent preference order.

If you have never encountered this paradox before [the so-called **paradox of majority rule**], you may be inclined to think that it is a trick or a scam or an illusion. Alas, not so. It is just as it appears. Majority Rule really is capable of generating an inconsistent social preference ordering. All of this was well known in the eighteenth century, and was, as we shall see later on, the subject of some imaginative elaboration by none other than the Reverend Dodgson, better known as Lewis Carroll. Enter now the young, brilliant economist Kenneth Arrow in the middle of the twentieth century. Coming out of a tradition of economic theorizing called Social Welfare Economics, to which a number of major figures, such as Abram Bergson, had contributed, Arrow conceived the idea of analyzing the underlying structure of the old Paradox of Majority Rule and generalizing it. The result, which he presented in his doctoral dissertation no less, was The General Possibility Theorem. Arrow published the theorem in 1951 in a monograph entitled *Social Choice and Individual Values*.

Another great economist and fellow Nobel Prize winner, Amartya Sen, in 1970 published *Collective Choice and Social Welfare*, in which he generalized and extended Arrow's work in astonishing ways. Sen's book is

difficult, but it is simply beautiful, and deeply satisfying. I strongly urge you, if you have a taste for this sort of thing, to tackle it. Sen has written widely and brilliantly on a host of extremely important social problems, including economic inequality, famine, and the demographic imbalance between men and women in the People's Republic of China. His little series of Radcliffe Lectures, published in 1973 as *On Economic Inequality*, is the finest use of formal methods to illuminate and analyze a social problem of which I am aware. It is a perfect example of the proper use of formal methods in social philosophy, and as such deserves your attention.

In *Collective Choice and Social Welfare*, Sen gives a simpler and more elegant proof of Arrow's General Possibility Theorem. Nevertheless, I have chosen in this blog to expound Arrow's original proof. Let me explain why. It often happens that the first appearance of an important new theorem is somewhat clumsy, valid no doubt, but longer and more complicated than necessary. Later theorists refine it and simplify it until what took many pages can be demonstrated quickly in a few lines. Sometimes, this development is unambiguously better, but at other times, the original proof, clumsy though it may be, reveals the central idea more perspicuously than the later simplifications do. I find this to be true in the case of Arrow's theorem. Sen's simplification serves several purposes, not the least of which is to set things up formally for his extremely important extension and elaboration of Arrow's work. Therefore, I urge you to look at it, once you have worked with me through Arrow's original proof.

Now let us begin. This is going to take a while, so settle down. Before we get into the weeds, let me try to explain in general terms what Arrow is doing. He asks, in effect, what are the underlying general assumptions of majoritarian decision making? What is it about voting with majority rule that appeals to us? He identifies five conditions or presuppositions [later reduced and simplified to four] that capture the logic of majority rule in a general way, and then shows that no way of making collective decisions that satisfies all four of them guarantees that the resulting social or collective choice will be consistent. This way of thinking about the problem accomplishes three things simultaneously. First, it unpacks majority rule voting into its component parts so that we can look at it and understand it

better. Second, it generalizes the Paradox of Majority Rule so that we realize we cannot avoid it simply by tweaking Majority Rule a bit [for example by requiring a two-thirds majority.] And finally, it allows us to see just exactly what Majority Rule does not do—in other words, it gives us insight into what would be totally different ways of making collective decisions.

We start with a series of assumptions, definitions, and notational conventions, some of which are already familiar to you from the opening segments of this general tutorial. This is going to be tedious, but learning these up now will make it infinitely easier to follow the proof. Here they are:

1. We start with a set of mutually exclusive alternatives, x, y, z, ... These may be all of the possible candidates in an election [i,.e., every single person who is eligible to hold office under the rules governing the election], every possible tax scheme that might come before Congress, all of the various possible decisions a City Council might take concerning zoning regulations, and so forth. The point of the phrase "mutually exclusive" is to rule out, for example, "Obama" and "Obama or Clinton" as two of the available alternatives.

2. On any give occasion when a decision is to be made, there is a subset, S, of the available alternatives, which will be called The Environment. This might be, for example, the relatively small number of people who have stated publicly that they would like to be elected to that office, or all the people who have formed campaign committees, or all the people who survive the primary season and are on the final ballot. Each of these is a subset of all the people eligible to hold the office [not necessarily a proper subset—i.e., not necessarily smaller than the total set of alternatives. All that is required is that S be included in the set of all alternatives, not that it be smaller than that set].

3. There is a set of individuals ["voters"], identified by numerical subscripts, 1, 2, 3, 4, ...

4. Each individual is assumed to have a **complete, transitive ranking**

of the entire set of alternatives, which we indicate using the notation introduced earlier—the binary relations R, I, and P. Just to review, xR_iy means that individual i considers alternative x to be as good as or better than alternative y. xP_iy and xI_iy are derived from R in the way indicated in the opening segments of this tutorial. What we are aiming for, of course, is a collective or social ranking, and that is indicated by the same letters, R, P, and I without the subscripts. So xPy means that the society prefers s to y. The whole point of this exercise is to start with complete, transitive individual rankings of the alternatives and then see whether there is any way of going from the individual rankings to a social ranking that satisfies certain conditions [see below] and results in a social ranking that is complete and transitive.

5. R_i all by itself refers to individual i's ranking of the entire set of alternatives, x, y, z,... Correspondingly, R all by itself refers to the society's ranking of the entire set of alternatives.

6. We shall have occasion to refer to different possible rankings, by an individual i, of the set of alternatives. We will indicate these different rankings by superscripts. So, for example, R_i is one ranking by individual i of the entire set of alternatives. R_i' is a second ranking. R_i'' is a third ranking. And R_i* is a fourth ranking. A ranking R_i can be thought of either as a list showing the way individual i ranks the alternatives, including ties [indifference], or as a set of all the ordered pairs (x, y) such that xR_iy.

7. A Social Welfare Function [an SWF] is a function that maps sets of individual rankings onto a social ranking. Such a mapping function qualifies as an SWF just in case both the individual rankings, the R_i, and the social or collective ranking, R, satisfy Axioms I and II below—which is to say, just in case the rankings, both individual and social, are complete and transitive.

8. A Social Welfare Function is said to be Dictatorial if there is some individual i such that, for all x and y, xP_iy implies xPy regardless of the

orderings of all of the individuals other than i. Thus, in particular, to say that an SWF is dictatorial is to say that there is some individual who can impose his or her will on the society with regard to the choice between any pair, x and y, even if everyone else in the society has the opposite preference as between those two alternatives.

9. Finally, we define something called a Social Choice Function [symbolized as $C(S)$.] $C(S)$ is the set of all alternatives in the Environment S such that, for every x and y in S, xRy. In other words, $C(S)$ is the set of top alternatives or best alternatives in S. Quite often, $C(S)$ will contain only one alternative, the one that the society prefers over all the others. But it may include more than one if the society is indifferent as among several best alternatives.

Those are the nine definitions and stipulations. The key new ones that we have not met before are S, the set of available alternatives, R, the social ranking, SWF, a Social Welfare Function, and $C(S)$, the Choice Function. Now Arrow lays down two Axioms governing the social ordering, R. These are:

Axiom 1 *For all x and y, xRy or yRx [Completeness]*

Axiom 2 *For all x, y, and z, if xRy and yRz then xRz. [Transitivity]*

O.K. So much for the preliminary throat clearing. I want you to go over these definitions and stipulations until you are comfortable with them. The proof is going to be a formal argument couched in terms of these symbols and appealing to these assumptions and axioms. You will find it impossible to follow if you do not have a solid grasp on these preliminary definitions and so forth. While you are doing that, I want to talk for a bit about several important points that are implicit in what we have just laid down, but may not be obvious.

From here on, I am going to break the exposition into short bits, because this is hard, and I do not want to lose anyone. My apologies to those of you who are having no trouble following it.

First of all, notice that Arrow assumes only ordinal preference. This means that there is no way in the proof to take account of intensity of preference, only order of preference. Let me give an example to make this clearer. In 1992, George H. W. Bush, Bill Clinton, and H. Ross Perot ran for the Presidency. There were some devoted followers of Perot who were crazy about him, and almost indifferent between Bush and Clinton, whom they viewed as both beltway politicians. Let us suppose that one of these supporters ranked Perot first, way ahead of the other two, and gave the edge slightly to Bush over Clinton, perhaps because Bush was a Republican. A second Perot supporter might have been rather unhappy with the choices offered that year, but preferred Perot slightly over Bush, while hating Clinton passionately. From Arrow's perspective which is that of ordinal preference, these two voters had identical preferences, namely Perot $>$ Bush $>$ Clinton, and an Arrovian SWF would treat the two individual preference orders as interchangeable.

Now, there are many ways in which citizens in America can give expression to the intensity of their preferences, as political scientists are fond of pointing out. One is simply by bothering to vote. Voter enthusiasm, in a nation half of whose eligible voters routinely fail to go to the polls, is a major determinant of the outcome of elections. A second way is by contributing to campaigns, volunteering for campaign work, and so forth. Yet another way is through a vast array of voluntary organizations dedicated to pursuing some issue agenda or advantaging some economic or regional group. *None of this can find expression in the sort of Social Welfare Function Arrow has defined*. This is a very important limitation on the method of collective decision that we call voting. Now, there are voting schemes that allow voters to give expression to the intensity of their preferences [such as giving each voter a number of votes, which he or she can spread around among many candidates or concentrate entirely on one candidate], but these too are ruled out by Arrow, who only allows the SWF to take account of individual ordinal preferences.

The second thing to note is that the requirement of completeness placed upon the SWF rules out partial orderings, such as those established by Pareto-Preference. It is often the case that every individual in

the society prefers some alternative x to some other alternative y, and if there are a number of such cases, a robust partial ordering might be established that, while not complete, nevertheless allows the society to rank a sizeable number of the available alternatives. This option too is ruled out by Arrow's two axioms. These observations have the virtue of helping us to understand just how restricting a collective decision-making apparatus like majority rule is. We are now ready to state the four conditions that Arrow defines as somehow capturing the spirit of majoritarian democracy. Arrow's theorem will simply be the proposition that there is no Social Welfare Function, defined as he has in the materials above, which is compatible with all four conditions. In the original form of the proof, the conditions were, as you might expect, called Conditions 1, 2, 3, 4, and 5. In the revised version, which I shall be setting forth here, they are called Conditions 1′ [a revised version of Condition 1], Condition 3 [which also is sometimes called the Independence of Irrelevant Alternatives], Condition P [for Pareto], and Condition 5. Here they are. I will tell you now that Condition 3 is the kinky one.

Condition 1′ All logically possible rankings of the alternative social states are permitted. This is a really interesting condition. What it says, formally speaking, is that each individual may order the alternatives, x, y, z, ... in any consistent way. What it rules out, not so obviously, is any religious or cultural or other constraint on preference. For example, if among the alternatives are various dietary rules, or rules governing abortions, or rules governing dress, nothing is ruled in or ruled out. The individuals are free to rank alternatives in any consistent manner.

Condition 3 Let R_1, R_2, ..., R_n and R'_1, R'_2, ... R'_n be two sets of individual orderings of the entire set of alternatives x, y, z, ...and let $C(S)$ and $C'(S)$ be the corresponding social choice functions. If, for all individuals i and all alternatives x and y in a given environment S, xR_iy if and only if xR'_iy, then $C(S)$ and $C'(S)$ are the same.

OK, this is confusing, so let us go through it slowly step by step and figure out what it means. To get to the punch line first, this condition says

that the society's eventual identification of best elements in an environment is going to be determined solely by the rankings by the individuals of the alternatives in that environment, and not by the rankings by the individuals of alternatives not in the environment. [Remember, the Environment, S, is a subset of all the possible alternatives.] Now, take the condition one phrase at a time. First of all, suppose we have two different sets of individual rankings of all the alternatives. The first set of rankings is the Ri [there are as many rankings in the set as there are individuals—namely, the first individual's ranking, R_1, the second individual's ranking, R_2, and so forth.] The second set of rankings is the R_i', which may be different from the first set.

Now, separate out some subset of alternatives, which we will call the Environment S, and focusing only on the alternatives in S, take a look at the way in which the individuals rank those alternatives, ignoring how they rank any of the alternatives left out of S. If the two sets of individual orderings, R_i and R_i', are exactly the same for the alternatives in S, then when the Social Welfare Function cranks out a social ranking, R, based on the individual orderings R_i and a social ranking, R', based on the individual orderings R_i', Condition 3 stipulates that the set of best elements [The Social Choice set] will be the same for R and for R'.

Whew, that still isn't very clear, is it? So let us ask the obvious question: What would this Condition rule out? Here is the answer, in the form of an elaborate example. Just follow along.

Suppose that in the 1992 presidential election, there are just three voters, whom we shall call 1, 2, and 3. Also, suppose there are a total of four eligible candidates: George H. W. Bush, Bill Clinton, H. Ross Perot, and me. Now suppose there are two alternative sets of the rankings of these four candidates by individuals 1, 2, and 3.

R_i: Individual 1: Wolff > Clinton > Bush > Perot
Individual 2: Bush > Perot > Wolff > Clinton
Individual 3: Wolff > Clinton > Bush > Perot

R_i': Individual 1: Clinton > Bush > Perot > Wolff
Individual 2: Bush > Perot > Clinton > Wolff
Individual 3: Clinton > Bush > Perot > Wolff

The crucial thing to notice about these two alternative sets of rankings is that they are identical with regard to the environment $S = (Bush, Clinton, Perot)$. The only difference between the two sets is that in the second set, Wolff has been moved to the bottom of everyone's list. [The voters find out I am an anarchist.]

Now let us consider the following Social Welfare Function: For each individual ranking, assign 10 points to the first choice, 7 points to the second choice, 3 points to the third choice, and 2 points to the fourth choice. Then, for any Environment, S, selected from the totality of available alternatives, determine the social ranking by adding up all of the points awarded to each alternative by the individual rankings. Got it?

Go ahead and carry out that exercise. If you do, you will find that for the first set of rankings, the Ri, and for the Environment $S = (Bush, Clinton, Perot)$. the SWF gives 16 points to Clinton, 16 points to Bush, and 11 points to Perot. So, $C(S)$, the society's decision as to which candidates are at the top, is (Clinton, Bush), because they each have the same number of points, namely 16. But if you now carry out the same process with regard to the second set of individual rankings, the R_i', and the same Environment S, you will discover that the SWF assigns 23 points to Clinton, 24 points to Bush, and 13 points to Perot, which means that $C'(S)$ is (Bush). So the social choice in the Environment S has changed, despite the fact that the relative rankings of the elements in S have not changed, because of a change in the rankings of an element not in S, namely Wolff. And this is just what Condition 3 rules out. It says that the Social Welfare Function cannot be one that could produce a result like this.

All of us are familiar with this sort of problem from sports meets or the Olympics. When we are trying to decide which team or country has done best, we have to find some way to add up Gold medals and Silver medal and Bronze medals, and maybe fourth and fifth places as well. And, as we all know, you get different results, depending on how many points you award for each type of medal. Arrow's Condition 3 rules out SWFs like that. Condition P: If xP_iy for all i, then xPy. This just says that if everyone strongly prefers x to y, so does the society. This is a very weak constraint

on the SWF.

Condition 5 The Social Welfare Function is not dictatorial. Remember the definition of "dictatorial" above. This rules out "*l'état c'est moi*" as a Social Welfare Function.

So, we have the definitions, etc., and we have the four Conditions that Arrow imposes on a Social Welfare Function. Remember that a Social Welfare Function is *defined* as a mapping that produces a social ranking that satisfies Axioms 1 and 2. Now Arrow is ready to state his theorem. It is quite simple:

There is no Social Welfare Function that satisfies the four Conditions.

This is really a devastating theorem. Basically, it says that there is no voting mechanism that gets around the Paradox of Majority Rule. The proof proceeds as follows. First Arrow states a set of little results about the relations R, I, and P. You are already familiar with them. They are trivial, as we shall see. Then he proves a little Lemma about the choice function. Then he proves a big important Lemma that is really the guts of the theorem. Finally, he uses the Lemmas to prove what is essentially an extension of the Paradox of Majority Rule, and he is done. We are going to go through this slowly and carefully. Let us start with the two little lemmas. Lemma 1 and Lemmas 2.

Lemma 1 1. *For all x, xRx*

2. *If xPy then xRy*

3. *If xPy and yPz then xPz*

4. *If xIy and yIz then xIz*

5. *For all x and y, either xRy or yPx*

6. *If xPy and yRz then xPz*

   These all follow immediately from the definitions of R, I, and P, the assumptions of transitivity and completeness, and truth functional logic. Arrow

includes them as an omnibus Lemma because at one point or another in his proof he will appeal to one or another of them. You should work through all the little proofs as an exercise. I will go through just one to show you what they look like.

7. *xRy or yRx [completeness]*

So if not xRy, then yRx. But the definition of yPx is yRx and not xRy Therefore, either xRy or yPx

Lemma 2 *xPy if and only if x is the sole element of $C([x,y])$*

If you review the definition of the Choice set, you will see that this Lemma is intuitively obvious. It says that in the little environment, S, consisting of nothing but x and y, if xPy, then x is the only element in the Choice set, $C(S)$. Since this is a bi-conditional [if and only if], we have to prove it in each direction.

1. Assume xPy. Then xRy, by Lemma 1(2). [See, this is why he put those little things in Lemma 1]. Furthermore, xRx, by Lemma 1(1). So x is in $C([x,y])$, because it is at least as good [i.e., R] as each of the elements of S, namely x and y. But if xPy then not yRx. Therefore, y is not in $C([x,y])$. So x is the sole element of $C([x,y])$.

2. Assume x is the sole element of $C([x,y])$. Since y is not in $C([x,y])$, not yRx. Therefore xPy.

Lemma 3 *If an individual, i, is decisive for some ordered pair (x,y) then i is a dictator.*

This is a rather surprising and very important Lemma. It is the key to the proof of Arrow's theorem, and shows us just how powerful the apparently innocuous Four Conditions really are. To understand the Lemma, you must first know what is meant by an ordered pair and then you must be given three definitions, including one for the notion of "decisive." Easy stuff first. An ordered pair is a pair in a specified order. An ordered pair is indicated by curved parentheses. Thus, the ordered pair (x,y) is the pair

$[x, y]$ in the order first x then y. As we shall see, to say that individual is decisive for some ordered pair (x, y) is to say that i can, speaking informally, make the society choose x over y regardless of what anyone else thinks. But a person might be decisive for x over y and not be decisive for y over x. We shall see in a moment how all this works out. Now let us turn to the three definitions that Arrow is going to make use of in the proof of Lemma 3.

Definition 1 *"A set of individuals V is decisive for* (x, y)*" =* df *" if* xP_iy *for all i in V and* yP_jx *for all j not in V, then* xPy

In other words, to say that a set of individuals V is decisive for the ordered pair (x,y) is to say that if everyone in V strongly prefers x to y, and everyone not in V strongly prefers y to x, then the society will strongly prefer x to y. Under majority rule, for example, any set of individuals V that has at least one more than half of all the individuals in the society in it is decisive for every ordered pair of alternatives (x, y).

Definition 2 *"*$x\underline{D}y$ *for i" or "i dictates over* (x, y)*" =* df *"If* xP_iy *then* xPy*"*

In words, we say that individual i dictates over the ordered pair (x,y) if whenever individual i strongly prefers x to y, so does the society regardless of how everyone else ranks x and y. [Notice that the capital letter D has a little line underneath it.]

Definition 3 *"*xDy *for i" or "i is decisive for* (x, y)*" =* df *"If* xP_iy*, and for all j not equal to i,* yP_jx*, then* xPy*."*

In words, i is said to be decisive for the ordered pair (x, y) if when i strongly prefers x to y and everyone else strongly prefers y to x, the society prefers x to y. [Notice that in this definition, the capital letter D does not have a little line underneath it.]

Ok. Now we are ready to state and prove the crucial Lemma 3.

Lemma 4 *Lemma 3: If* xDy *for i, the* $z\underline{D}w$ *for i, for all z,w in S*

In words, what this says is that if any individual, i, is decisive for some ordered pair (x, y) then that individual i is a dictator [i.e., dictates over any ordered pair (z, w) chosen from S]. This is an astonishing result. It says that if the Social Welfare Function allows someone to compel the society to follow her ranking of some ordered pair, no matter what, against the opposition of everyone else, then the Social Welfare Function makes her an absolute dictator. [*L'ėtat c'est moi*]. Here is the proof. It is going to take a while, so settle down. In order to make this manageable, I must use the various symbols we have defined. Let me review them here, so that I do not need to keep repeating myself.

An ordered pair is indicated by curved parentheses: (x, y), as opposed to a non-ordered pair, which is indicated by brackets: $[x, y]$.

$x\underline{D}y$ for i, which is D with a line under it, means "i dictates over (x, y)" (an ordered pair) xDy for i, which is D with no line under it, means "i is decisive for (x, y)"

Proof of Lemma 3 Assume xDy for i [i,e., i is decisive for x against y]

The proof now proceeds in two stages. First, for an environment $[x, y, z]$, constructed by adding some randomly chosen third element z to x and y, we show that i is a dictator over $[x, y, z]$.

Then we show how to extend this result step by step to the conclusion that i is a dictator over the entire environment S of admissible alternatives.

First Stage Proof that i is a dictator over the environment $[x, y, z]$

(step i) Construct a set of individual orderings over $[x, y, z]$ as follows.

R_i: x > y > z [i.e., individual i's ordering of the three]

All the other R_j: yP_jx $yP_jzR_j[x, z]$ unspecified

In other words, we will prove something that is true regardless of how everyone other than i ranks x against z.

(step ii) xP_iy by construction. But, by hypothesis xDy for i. Therefore xPy

In words, i is assumed to strongly prefer x to y, and since by hypothesis i is decisive for x against y, the society also strongly prefers x to y.

(step iii) For all i, yP_iz, by construction. Therefore, yPz, by Condition P, and xPz by Lemma 1(3). In words, since everyone strongly prefers y to z, so does the society. And since the society strongly prefers x to y and y to z, it strongly prefers x to z [since Axiom 2, which is used to prove Lemma 1(3), stipulates that the SWF is transitive.]

(step iv) So xPz when xP_iz, regardless of how anyone else ranks x and z. [check the construction of the individual orderings in step (i)]

(step v) Hence $x\underline{D}z$ for i, which is to say that i dictates over the ordered pair (x, z)

(step vi) Now consider (y, z) and assume the following set of individual orderings:

$R_i : y > x > z$ All the other $R_j : yP_jx\ zP_jx$ and $R_j[y, z]$ unspecified.

(step vii) yP_ix for all i. Therefore yPx by Condition P

(step viii) $x\underline{D}z$ for i, by (v). Hence xPz.

(step ix) So yPz by Lemma 1(3). Thus $y\underline{D}z$ for i.

In words, we have now shown that i dictates over the ordered pair (y, z). Let us take a minute to review what is going on here. We are trying to prove that if i is decisive for a single ordered pair, (x, y), then i is a dictator over an environment consisting of x, y, and some randomly chosen z. If we can show that i is a dictator for every ordered pair in the environment $[x, y, z]$ then we shall have shown that i is a dictator over that environment. There are six ordered pairs that can be selected from the environment, namely (x, y), (x, z), (y, x), (y, z), (z, x), and (z, y). So we must establish that i dictates over every single one of these ordered pairs. We have already established that i dictates over (x, z) in step (v) and over (y, z) in step (ix).

(step x) We can now extend this argument to the other four ordered pairs that can be selected from the environment $[x, y, z]$. In particular, let us do this for the ordered pair (y, x). Construct the following set of orderings:

$R_i : y > z > x$

All the other $R_j : zP_jy\ zP_jx\ R_j[x, y]$ unspecified.

(step xi) zP_ix for all i. Hence zPx by Condition P

(step xii) $y\underline{D}z$ for i by (step ix). Hence yPz

(step xiii) So yPx by Lemma 1(3). Thus $y\underline{D}x$ for i.

So we have proved [or can do so, by just iterating these steps a few more times] that i dictates over every ordered pair in $[x, y, z]$, and therefore i is a dictator over the environment $[x, y, z]$. So much for Stage One of the proof of Lemma 3. Now, take a deep breath, review what has just happened to make sure you understand it, and we will continue to:

Stage Two The extension of our result to the entire environment, S, of available alternatives. Keep in mind that S, however large it may be, is finite. Assume xDy for i [our initial assumption—just repeating for clarity] and also assume the result of Stage One. Now consider any ordered pair of alternatives (z, w) selected from the environment S. There are just seven possibilities.

1. $x = z$, w is a third alternative
2. $x = w$, z is a third alternative
3. $y = z$, w is a third alternative
4. $y = w$, z is a third alternative
5. $x = z$, $y = w$
6. $y = zx = w$
7. Neither z nor w is either x or y

Case 1: We have an environment consisting of three alternatives: $[x = z, y, w]$. Stage One shows that if xDy for i, then $x = z\underline{D}w$ for i.
Case 2, 3, 4: Similarly
Case 5: Trivial
Case 6: Add any other element v to form the environment $[x = w, y = z, v]$. From $x = wDy = z$ for i, it follows that $y = z\underline{D}x = w$ for i. [In words, just in case you are getting lost: In the case in which y is element z and x is element w, from the fact that i is decisive for x against y, which is to say

for w against z, it follows that i dictates over y and x, which is to say over z and w. This is just a recap of Stage One.

Case 7: This is the only potentially problematic one case, and it needs a little explaining. We are starting from the assumption that i is decisive for x against y, and we want to show that i is a dictator over some totally different of alternatives z and w, so we are going to creep up on that conclusion, as it were. First we will add one of those two other alternatives, z, to the two alternatives x and y to form the environment $[x, y, z]$. From Stage One, if xDy for i then $x\underline{D}z$ for i. But trivially, since $x\underline{D}z$ for i, it follows that xDz for i. [The point is that if i dictates over x and z, then of course i is decisive for x against z].

Now add w to x and z to form the environment $[x, z, w]$. Since xDz for i, it follows that $z\underline{D}w$ for i, by Stage One. In words, if i is decisive for x against z, then in the environment $[x, z, w]$, i dictates over z and w. This follows from Stage One. What this shows is just how powerful Lemma 3 really is.

Thus we have demonstrated that xDy for i implies $z\underline{D}w$ for i, for all z and w in S. In other words, if i is decisive for some ordered pair (x, y), then i is a dictator over S. But Condition 5 stipulates that no individual may be a dictator.

Therefore:

An acceptable Social Welfare Function does not permit any individual to be decisive for even a single ordered pair of alternatives in the environment S of available alternatives.

Can we all say Ta-Da? This is the heavy lifting in Arrow's theorem. Using this Lemma, we can now fairly quickly prove that there is no SWF satisfying Axioms 1 and 2 and all four Conditions, 1', 3, P, and D.

4.3.1 Proof of Arrow's Theorem

Step 1. By Condition P, there is at least one decisive set for each ordered pair, namely the set of all the individuals. From all the decisive sets, choose a smallest decisive set, V, and let it be decisive for some ordered pair (x, y).

What I mean is this: Consider each set of individuals that is decisive for

some ordered pair or other. Since there is a finite number of individuals, each of these sets must have some finite number of individuals in it. And the sets may have very different numbers of individuals in them. But one or more of them must be the smallest set. So arbitrarily choose one of the smallest, call it V, and label the pair of alternatives over which it is decisive (x, y).

Step 2: By Condition P, V cannot be empty. [Go back and look at Condition P and make sure you see why this is so. It is not hard]. Furthermore, by Lemma 3, V cannot have only one member [because Lemma 3 proved that no single individual, i, can be decisive for any ordered pair (x, y)]. Therefore, V must have at least two members.

Step 3: Partition the individuals $1, 2, \ldots, n$ in the following way: The set of all individuals

$$\begin{array}{|c|c|}\hline V & V_3 \\ \hline V_1 & V_2 \\ \hline \end{array}$$

Where,

V_1 = a set containing exactly one individual in V

V_2 = the set of all members of V except the one individual in V_1

V_3 = the rest of the individuals, if there are any.

Is this clear? V is a smallest decisive set. It must have at least two individuals in it. So it can be divided into V_1 containing just one individual, and V_2 containing the rest of V. V_3 is then everyone else, if there is anyone else not in the smallest decisive set V.

Step 4: Now let the individuals in the society have the following rankings of three alternatives, x, y, and z. [And now you will see how this is an extension of the original Paradox of Majority Rule with which we

began.]

$V_1 : x > y$ and $y > z$
$V_2 : z > x$ and $x > y$
$V_3 : y > z$ and $z > x$

[You see? This is one of those circular sets of preference orders: xyz, zxy, yzx]

V_1 is non-empty, by construction.
V_2 is non-empty, by the previous argument.
V_3 may be empty.

Step 5:

1. By hypothesis, V is decisive for x against y. But V is the union of V_1 and V_2, and xP_iy for all i in V_1 and V_2. Therefore, xPy. [i.e., the society prefers x to y.]

2. For all i in the union of V_1 and V_3, yP_iz. For all j in V_2, zP_jy. If zPy, then V_2 is decisive for (x, y). But by construction, V_2 is too small to be decisive for anything against anything, because V_2 is one individual smaller than a smallest decisive set, V. Therefore not zPy. Hence, yRz [see the definitions of P and R].

3. Therefore xPz by Lemma 1(6) [go back and look at it].

4. But xP_1z and zP_ix for all i not in V_1, so it cannot be that xPz, because that would make V_1 decisive for (x, z), which contradicts Lemma 3. Therefore, not xPz.

Step 6: The conclusion of Step 5.4 contradicts Step 5.3. Thus, we have derived a contradiction from the assumption that there is a Social Welfare Function that satisfies Conditions 1′, 3, P, and 5. Therefore, there is no

SWF that satisfies the four Conditions. *Quod erat demonstrandum.*

OK. Everybody, take a deep breath. This is a lot to absorb. Arrow's Theorem is a major result, and it deserves to be studied carefully. Go back and re-read what I have written and make sure you understand every step. It is not obscure. It is just a little complicated. If you have questions, post them as a comment to this blog and I will answer them.

With this segment, I conclude my discussion of Arrow's General Possibility Theorem. And I think this will also conclude this tutorial on the use and abuse of formal methods in political philosophy. I will be happy to respond to questions, if there are any, but I think enough is enough. Thank you all for staying with me on this, for pointing out errors, and for asking questions. It has been fun for me, revisiting material I have not taught for twenty years or more, and I hope it has been informative and fun for you.

An extremely interesting result concerning the consistency of majority rule was produced by the Australian political scientist Duncan Black. In a book called *The Theory of Committees and Elections* (Black, 1958), published in 1958, Black proved an important theorem about circumstances under which majority rule is guaranteed to produce a transitive social preference ordering. In a moment, I am going to go through the proof in detail, but let me first explain intuitively what Black proved. Ever since the French Revolution, political commentators have adopted a convention derived from the seating arrangement in the National Assembly. In that body, Representatives belonging to each party were seated together, and the groups were arrayed in the meeting hall in such a manner that the most radical party, the Jacobins, sat on the extreme left of the hall, and the most reactionary party, the Monarchists, sat on the right, with the other groups seated between them from left to right according to the degree to which their policies deviated from one extreme or the other. Thus was born the left-right political spectrum with which we are all familiar. [Of course, in the U. S. Senate, there are no Communists and only one Socialist, but, as the reign of George W. Bush shows, there are still plenty of Monarchists.]

The interesting fact, crucial for Black's proof, is that wherever a party

locates itself on the spectrum, it tends to prefer the positions of the other parties, either to the left or to the right, less and less the farther away they are seated. So, if an individual identifies himself with a party in the middle, he will prefer that party's positions to those of a party a little bit to the left, and he will prefer the policies of the party a little bit to the left over those of a party farther to the left, and so on. The same is true looking to the right. Notice that since only ordinal preference is assumed, you cannot ask, "Is a party somewhat to the left of you farther from you than a party somewhat to the right of you?" [Make sure you understand why this is true. Ask me if it is not.]

Consider contemporary American politics. If I am a moderate Republican [assuming there still is one], I will prefer my position to that of a conservative Republican, and I will prefer that position to a right wing nut. I will also prefer my position to that of a Blue Dog Democrat [looking to my left rather than to my right], and that position to the position of a Liberal Democrat, and that position in turn to the position of a Socialist [Bernie Sanders?].

This can be summarized very nicely on a graph, along the X-axis of which you lay out the left-right political spectrum, while on the Y-axis you represent the order of your preference. Pretty obviously, the graph you draw will have a single peak—namely, where your first choice is on the X-axis—and will fall away on each side, going monotonically lower the farther you get on the X-axis from your location on it. In short, your preference, when graphed in this manner, will be single-peaked. Here is an example of a person's preference order graphed in this manner. For purposes of this example, there are five alternatives, (a, b, c, d, e), and the individual has the following preference order: $d > e > c > b > a$ (see Figure 4.1).

Let us suppose that there is a second person whose preference order is $a > b > c > d > e$. It is obvious that if we posted this person's preferences on the same graph, the two together would look like Figure 4.2.

Notice that each of these lines has a single peak. The first individual's

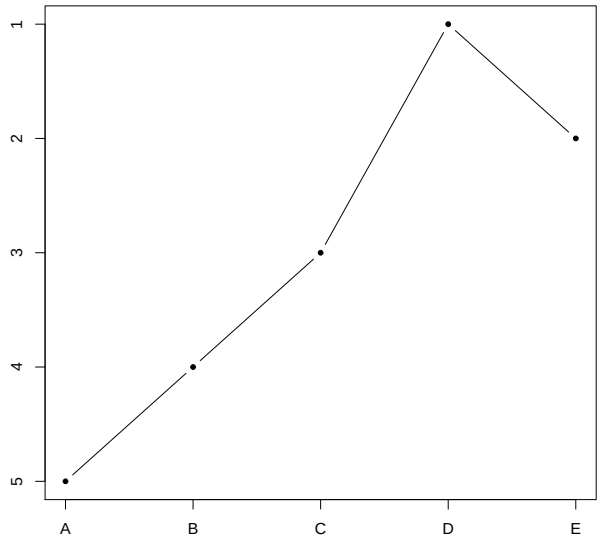


Figure 4.1: Preference ordering of an individual.

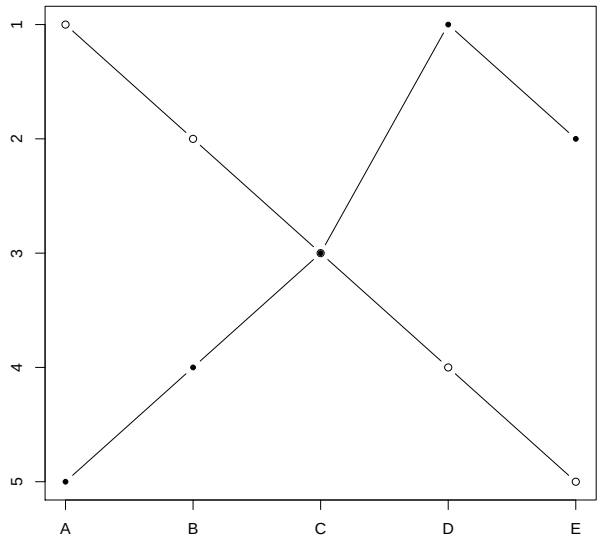


Figure 4.2: Preference ordering of two people.

line peaks at alternative D; the second's at alternative A. If you do a little experimenting, you will find that if you change the order in which the alternatives are laid out on the X-axis, sometimes both lines are still single peaked, sometimes one remains single peaked and one no longer is. Sometimes neither is single peaked. For example, if you change the order slightly so that the alternatives are laid out on the X-axis in the order a B E D C, the first individual's line will still be single peaked, but the second individual's line will now be in the shape of a V with one peak at A and another peak at C. [Try it and see. It is too much trouble for me to draw it and scan it and size it and insert it.]

Suppose now that we have an entire voting population, each with his or her own preference order, and that we plot all of those preference orders on a single graph, a separate line for each person. There might be some way of arranging the alternatives along the X-axis so that everyone's preference order, when plotted on that graph, is single peaked. Then again, there might not be. For example, if you have three people and three alternatives, and if those three people have the preferences that give rise to the Paradox of Majority Rule, then there is no way of arranging the three alternatives along the X-axis so that all three individuals' preferences orders can be plotted on that graph single-peakedly. [Try it and see. Remember that mirror images are equivalent for these purposes, so there are really three possible ways of arranging the alternatives along the X-axis, namely xyz, xzy, and yxz.]

Duncan Black proved that if there is some way of arranging the available alternatives along the X-axis so that everyone's preference order, plotted on that graph, is single peaked, then majority rule is guaranteed to produce a consistent social preference order. Notice, in particular, that if everyone's preferences can be mapped onto the familiar left-right spectrum, with each individual preferring an alternative less and less the farther away it is in either direction from the most preferred alternative, then everyone will on that graph have a single peaked order [because it will peak at the most preferred alternative and fall away monotonically to the right and to the left.]

The proof is fairly simple. It goes like this.

Step 1: Assume that there are an odd number of individuals [the proof works for an even number of individuals, but in that case there can be ties, which produces social indifference, which then requires an extra couple of steps in the proof, so I am trying to make this as simple as possible.] Assume that their preferences can be plotted onto a graph so that all of the plots are single-peaked.

Step 2: Starting at the left, count peaks [there may be many peaks at the same point, of course, showing that all of those people ranked that alternative as first] and keep counting until you reach one more than half of the total number of peaks, i.e. $(n/2+1)$. Assume there are p peaks to the left of that point, q peaks at that point, and r peaks to the right, with $(p+q+r) = n$. Now, by construction, $(p+q) > n/2$ and $pn/2$, because if $(q+r)n/2$, which by construction it is not.

Step 3: Let us call the alternative with the q peaks alternative x. Clearly, there is a majority of individuals who prefer x to every alternative to the right of x on the graph, because there are $p+q$ individuals whose plots are downward sloping from x as you go to the right, which means they prefer x to everything to the right, and $p+q$ is a majority. But there are $q+r$ individuals who prefer x to everything to the left of x, because their plots are downward sloping as you go to the left, and $q+r$ are a majority. So alternative x is preferred in a pairwise comparison by a majority to every other alternative.

Step 4: Remove alternative x from the graph, remove alternative x from everyone's preference order, and then redraw all of the plots. They will all still be single-peaked. Why? Well, there are three possible cases: Either the dot representing the individual's ranking of x was the peak, or it was to the left of the peak, or it was to the right. In each case, when you reconnect the remaining dots, the graph remains single-peaked [try it and

see. It is too hard to draw it and scan it and upload it. But it is intuitively obvious.]

Step 5: You now have a new set of single-peaked plots on a single graph, so go through Steps 2 and 3 all over again. The winning alternative is preferred to every other remaining alternative, and is of course inferior to the first winner. If you now iterate this process until you run out of alternatives, you are left with a fully transitive social preference established by repeated uses of majority rule.

Black's theorem has considerable real world application, as we have seen, but it of course does not identify necessary and sufficient conditions for majority rule to produce a transitive social preference order. It only identifies a sufficient condition, namely single-peakedness. This means that there are sets of individual preferences that cannot be mapped single-peakedly onto a single graph, and yet which by majority rule produce transitive social preference orders. I leave it to you to construct an example of this.

References

Arrow, K. J. (1963). *Social Choice and Individual Values*. Cowles Foundation / John Wiley & Sons, New York, 2nd edition.

Black, D. (1958). *The Theory of Committees and Elections*. Cambridge University Press, Cambridge, MA.

Elster, J. (1985). *Making sense of Marx*. Cambridge University Press, Cambridge, MA.

Neumann, J. v. and Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Blackwell, Oxford, UK.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge (MA).

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA / London, UK.

Stevens, S. S., editor (1951). *Handbook of Experimental Psychology*. Wiley & Sons, New York, NY, 1st ed. edition.

Wolff, R. P. (1977a). A formal analysis of the bargaining game. In *Understanding Rawls: A Reconstruction and Critique of A Theory of Justice*, pages 142–179. Princeton University Press, Princeton, NJ.

Wolff, R. P. (1977b). Robert Nozick's derivation of the minimal state. *Arizona Law Review*, 79:7–30.

Wolff, R. P. (1990). Methodological individualism and Marx: some remarks on Jon Elster, game theory, and other things. *Canadian Journal of Philosophy*, 20(4):469–486.