2

Utility Theory

OVERVIEW

The analysis of economic rationality in Chapter 1 was somewhat unorthodox. Typically discussions of economic rationality commence with the idea of *preferences*, the axioms of *ordinal utility* theory, and then, finally, *cardinal utility* theory. Instead, I have tried to determine the relation of *Homo Economicus* to instrumental—goal-based—rationality. I have two reasons for adopting this slightly unorthodox approach. *First*, it is important to see how economic rationality relates to our broader, nontechnical conceptions of practical rationality. *Second*, as I will try to show in this chapter, the relation of formal utility to instrumental rationality, and so to *Homo Economicus*, is usually misunderstood. Now that we have a reasonably firm grasp of instrumental rationality, we are well-positioned to avoid some common errors.

Utility theory is about satisfying preferences. I begin by considering an often overlooked question: just what is a "preference"? The next two sections explain ordinal and cardinal utility theory. After explaining the basics of utility theory, I consider whether utility theory is, as many think, a formalization of the idea of instrumental rationality. I argue that it is not: it can model both instrumental and noninstrumental reasoning. The chapter concludes by discussing the experiments of social psychologists as to whether utility theory really explains human behavior.

2.1 PREFERENCES: WHAT ARE THEY?

The idea of a "preference"—and, especially, of "satisfying a preference"—is fundamental to utility theory. Economists explain rational action as that which aims at preference satisfaction: a rational agent is typically held to maximize the satisfaction of her preferences (see Section 1.3). Unfortunately, "preference" is an especially ambiguous term. We can identify at least three interpretations of preference: as (1) a noncomparative "taste" for something; (2) "choice behavior" itself; and (3) the agent's deliberative rankings of her outcomes and options.

Nonrelational Tastes or Desires

Philosophers, lawyers, and even some economists tend to equate the idea of a preference with a liking. To have a "preference for pizza" is to like pizza, or to have a taste for pizza.[1] "I prefer it" means "I like it" or "I have a taste for it." Thus, for Louis Kaplow and Steven Shavell, to say a person has a preference for a fair outcome is simply to say that she has a "taste" for fairness: "if individuals in fact have tastes for notions of fairness—that is, if they feel better off when laws or events that they observe are in accord with what they consider to be fair—then analysis under welfare economics will take such tastes into account. . . ."[2] Notice the echo of hedonism (Section 1.1): a person's sole aim seems to be to "*feel* better off."

Now although we sometimes talk this way, there are two reasons why this notion of preference cannot enter into utility theory. (1) Consider a famous story presented by Michael Walzer:

> a politician who has seized upon a national crisis—a prolonged colonial war—to reach for power. He and his friends win office pledged to decolonization and peace; they are honestly committed to both, though not without some sense of the advantages of the commitment. In any case, they have no responsibility for the war; they have steadfastly opposed it. Immediately, the politician goes off to the colonial capital to open negotiations with the rebels. But the capital is in the grip of a terrorist campaign and the first decision the new leader faces is this: he is asked to authorize the torture of a captured rebel who knows or probably knows the location of a number of bombs hidden in apartment buildings around

> the city, set to go off in the next twenty four hours. He orders the man tortured, convinced that he must do so for the sake of the people who might otherwise die in the explosions—even though he believes that torture is wrong, indeed abominable, not just sometimes, but always.[3]

If we think of a "preference" as something akin to "liking" or "having a taste for," we can interpret this case in two ways: we can say (a) that the politician has a preference (taste) for torture or (b) that he does *not have a preference* to torture the terrorist rather than to let the buildings blow up, though *he does choose* to torture the terrorist. The first interpretation is obviously wrong. As Walzer tells the story the politician despises torture; he certainly does not have a taste for it. So perhaps we should adopt the second interpretation, and say that the politician chooses to torture but does not have a preference for it (since he certainly does not like it). But this is inconsistent with utility theory. Under utility theory, if our politician ranks option x as more choice-worthy than y, then he prefers x, even if he detests both (though he detests x a little less than y).[4] If he is rational and he chooses x then he *must* prefer x to y, even though he doesn't like either. We need, then, to make sure that we do not confuse the technical notion of a preference with the ordinary language conception of a "liking."

(2) More importantly, understanding preferences as tastes is to understand preferences as noncomparative: I like x, or have a taste for x, or desire x, where x is one thing or option. But in utility theory preferences are always understood as comparative: one always prefers one thing (or option) to another. A preference always and necessarily relates two options and compares them in terms of choice-worthiness. In utility theory one simply cannot have a preference for one option. In this way "preference" is more like "bigger" than "big." One thing can be "big," but "bigger" relates two things: it is inherently comparative. We now see why it is a confusion to take "preference" as synonymous with "goals," "desires," or "values" (Section 1.1). The latter ideas are all noncomparative: my goal can be just x, I can desire just x, or I can simply value x. But in utility theory I cannot simply have a preference for x—it must be a preference for x over y, some second option. We must, then, *define* preference as the "$\succ$" relation, such that $x \succ y$ means x is preferred to y—it is more choice-worthy.

Revealed Preferences

Economists often insist—at least in their more official pronouncements—on a purely behavioral conception of preference: Alf is said to have a preference for *x* over *y* if and only if Alf chooses *x* over *y*, where choice is conceived of as overt behavior. On this view to prefer *x* to *y* is simply to choose it over *y*; if one has never made the choice then one does not have the preference. Preference so understood is, then, equivalent to actual choice. When pressed, economists are apt to say that a preference is simply choice behavior, and if one has consistent preferences this means simply that one chooses consistently. Thus, it is said, one's actual choices "reveal" one's preferences.

The very term "revealed" preference is somewhat misleading. If preferences just *are* choices, what sense can be made of saying that a choice *reveals* a preference? To use this sort of language is to suggest that the choice is "revealing" something else, something hidden and mental, as when a person makes a "revealing statement," showing something previously hidden about her character. However, avoiding any appeal to such mental entities was the explicit aim of the behavioristically inclined economists who stressed "revealed preference" theory.[5] Leaving aside the confusion about what is supposed to be "revealed," we have powerful reasons to question the plausibility of the behavioristic project; the attempt to rid the mental from social science looks doomed to failure. Choice is an intentional concept: any effort to describe Alf's choice of *x* over *y* will necessarily involve a reference to his understanding of what he is doing, and the nature of the choices confronting him (see further Section 2.3). "Voting for candidate *x* over *y*" for example, is not a piece of behavior qua movement of a body. A description of an act as "a vote for *x*" necessarily turns on the intentions—mental states—of actors involved. The behavior of "raising an arm" may be the act of asking a question or casting a vote (or innumerable other acts); only reference to the intentions of the agent can distinguish the two. And very different pieces of behavior (raising a hand, marking a piece of paper, or shouting "yea!") all may constitute the same act of "voting for *x*." This is not to say that an intention is sufficient (I cannot vote for the President of Russia even if I intend to), but it seems quite impossible to rid the intentional from our conception of choice. If so, we can hardly purge the mental from our explanation of choices.

CHAPTER 2

Deliberative Preferences

We cannot do without appeal to the mental in accounting for what is involved in choosing on the basis of one's preferences. Although in their official pronouncements economists are apt to adopt a behavioristic notion of revealed preference, most economic writing, and almost all accounts of rational action, suppose that actual choice is taken to *reveal* (or advance) preference qua a deliberative ranking of the options by the agent. A person deliberates and, ideally, can rank all the possible "outcomes"—the ways in which things may go or, as philosophers sometimes say, "states of affairs." Of course it is impossible to actually do this: you would have to identify every possible event that might result from your action. The number of states of affairs that could result from your action looks indefinitely large, if not infinite. I shall return in Section 2.4 to this fundamental problem of how to specify the options; for now I follow the common practice of assuming away the problem by stipulating some finite feasible set of mutually exclusive options that the agent orders in terms of choice-worthiness. The agent is assumed to order the options in the feasible set—her actual choice from the feasible set would then *reveal* her deliberative preference over the feasible set of options.

This leads us to a fundamental issue in utility theory: what ultimately do our preferences range over—states of affairs (outcomes) or possible actions that we might undertake? Say you prefer to live in New Orleans rather than New York (states of affairs), but now you have to consider how this preference over outcomes relates to what action you are to choose. Perhaps the actions open to you are to accept an offer of going to chef school in New Orleans and accepting an offer of law school in New York: if your deliberative preference over outcomes is to live in New Orleans, perhaps the action you should choose is to accept the chef school offer because that action-option will better satisfy your deliberative preference over outcomes. It looks like we must, then, consider both a person's ranking of *outcomes* (roughly, how much one values the various states of affairs that one can bring about) and a person's ranking of *action-options* in terms of which action is most choice-worthy.[6] Let us say that the *consequence domain* of a choice involves everything relevant to a person's ranking of states of the world that might obtain (the world in which you earn a million dollars a year through your legal practice, the world in which you earn half a million by being a successful chef, the world in which you eat great food, the world in which you eat

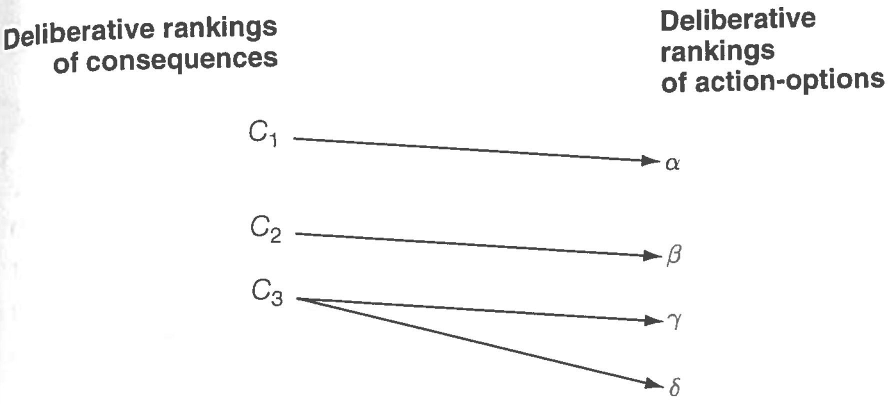
Deliberative rankings of consequences
Deliberative rankings of action-options
C_1
C_2
C_3
α
β
γ
δ

FIGURE 2-1 Mapping Actions to Outcomes

yogurt).[7] Suppose you can rank all these states of the world in terms of which best satisfies your goals and desires, etc. However, the important point for utility theory is not *why* you rank states of the world the way you do, but that you are *able* rank them. Now suppose that you confront a variety of action-options (a choice to be a lawyer or a chef); you will rank the actions highest that are correlated with the highest-ranked outcomes.[8] *Thus, your preferences over outcomes determine your preferences over options.*[9] We can think, then, of a *mapping* of an ordering of outcome-consequences on to the action-options, producing an ordering of action-options as in Figure 2-1.

Although the ordering of action-options (α, β . . .) will be correlated with the ordering of outcome-consequences ($C_1 \ldots C_n$), there need not be a unique one-to-one relation. As we see in Figure 2-1, two actions (γ, δ) may be correlated with the same consequence C_3: in this case a rational agent will be indifferent between γ and δ since they are associated with the same consequence. If, on the other hand, α is correlated with C_1 and β is correlated with C_2, where C_1 is a higher ranked consequence than C_2, then a rational agent will prefer α over β. We suppose that a person's preferences over outcomes are simply given;[10] one's preferences over action-options, however, change as one gets new information about the relation of the acts to one's rankings of outcomes. The distinction between preferences over outcomes and over actions is especially important when we get to utility theory under risk and strategic situations such as games, where one cannot be certain what action-options produce certain outcomes (see Section 2.3; Chapter 4).

2.2 ORDINAL UTILITY THEORY

The Axioms

Let us focus on the core idea of preferences over outcomes: they are clearly at the heart of the story (we will reintroduce the distinction between such preferences and preferences over options in the next section). We can generate an *ordinal utility function* for any person in terms of his preference rankings for the different outcomes if his rankings satisfy the following standard conditions for a *weak* ordering:

1. The ordering is *complete*. If Alf has a complete ordering, he can always rank options in his feasible set—he can always decide whether one possible outcome is better than another, worse than the other, or equally choice-worthy. Note that in this sense Alf "orders" a pizza and a taco if either (a) he strictly prefers one to the other or (b) he is indifferent between them: a pizza and a taco are equally worthy of being chosen. More formally, we can say that for every pair of outcomes (x,y) it must be the case that in Alf's ordering: (a) x is strictly preferred to y; or (b) y is strictly preferred to x; or (c) y and x are indifferent. Let us use "$x \succ y$" for "x is strictly preferred to y"; "$\mathrm{x} \sim y$" for "x is indifferent to y"; and "$x \succeq y$" for "x is either preferred to y or x is indifferent to y." So for all (x,y): $x \succeq y$ $\underline{\vee}$ [i.e., or] $y \succeq \mathrm{x}$.[11] I shall call "$x \succeq y$" the general preference relation, and "$x \succ y$" the strict preference relation.
2. If Alf strictly prefers a pizza to a taco, it must be the case that he does *not* strictly prefer a taco to a pizza. The strict preference relation is, then, *asymmetric*: $\neg$ [i.e., not] $(x \succ y\ \&\ y \succ \mathrm{x})$. In contrast, if Alf is indifferent between a pizza and a taco, he is also indifferent between a taco and a pizza; indifference, therefore, is *symmetric*: $(x \sim y)$ if and only if $(y \sim x)$.
3. We also need a rather obvious but uninteresting axiom: Alf must hold that a pizza is at least as good as a . . . pizza! The general preference relation is *reflexive*: $x \succeq x$.
4. More interestingly, Alf's preferences must be transitive. If Alf prefers a pizza (x) to a taco (y), and a taco (y) to a cup of yogurt (z), then Alf must prefer a pizza (x) to a cup of yogurt (z). Also, if Alf is indifferent between a pizza (x) and a taco (y), and indifferent between a taco (y) and a cup of

TABLE 2-1 Three Equivalent Ordinal Utility Functions

Preference	μ function A	μ function B	μ function C
x	3	10	1000
y	2	5	99
z	1	0	1

> yogurt (z) (!), then Alf must be indifferent between a pizza (x) and a cup of yogurt (z). So, more formally, $(x \succeq y)$ & $(y \succeq z) \rightarrow x \succeq z$. (Both strict preference and indifference are transitive.)

We can now define *utility* in terms of preference. Letting μ stand for utility, we can say that $x \succ y \equiv \mu(x) > \mu(y)$ (i.e., "x is strictly preferred to y" is equivalent to "the utility of x is greater than the utility of y"). It is, then, an error to say (as is all-too-often said) that a person prefers x to y *because* x gives him more utility. Utility does not explain the preference: *utility is simply a representation of the preference.* Utility is not something apart from, or additional to, preference satisfaction: it is a numerical function that represents the degree to which a person's preferences are satisfied. Ordinal utility functions map preferences over outcomes on to numbers. If we assume that the most preferred outcome is mapped on to the highest number, then the next preferred is mapped onto a smaller number, the next on to a yet smaller number, and so on. The sizes of the differences, or ratios between the numbers, provide no additional information. A person's preference ranking can generate an infinite number of ordinal utility measures: the strict preferences $x \succ y \succ z$ might be represented by any of the three utility functions in Table 2-1.

It should be clear that it makes no sense to add together different people's ordinal utilities (or even to add a single person's ordinal utilities for different outcomes). All the ordinal utility function tells us is that, for a specific person, a higher-numbered outcome is preferred to a lower-numbered one.

Why Accept the Axioms?

Can we show people that they *should* order outcomes according to the ordinal axioms? Suppose someone challenges the transitivity axiom:

> Yes, I can understand what transitivity is. According to transitivity, if I prefer x to y, and I prefer y to z, I must

> prefer x to z. But in fact that isn't the way my preferences go. I prefer a pizza (x) to a taco (y), and a taco (y) to a cup of yogurt (z), but I just do prefer a cup of yogurt (z) to a pizza (x)! Given any pair of options I can always make a choice. So what's wrong with that?[12]

Before we proceed, I want to point out that this is an extreme case: we assume that the person *simultaneously* asserts all three strict preferences ($x \succ y$, $y \succ z$, $z \succ x$). With actual agents, we would expect them to choose x over y at one point in time, y over z at another, and finally z over x at yet another. In these sorts of sequential choices we can only infer that the person's preferences violate transitivity if we assume that her preferences are *stable*. One possible explanation of her third choice (of z over x) is she has now tired of pizza; in that case she has undergone preference change and we cannot say that her preferences violate transitivity. If her preferences shift back and forth from moment to moment, we could never infer that her ordering violates transitivity. Thus we need either to suppose stable preferences over outcomes or a nice case in which at one moment the person entertains all three preferences.

Many respond to the above challenge by invoking once again the idea of instrumental rationality (Section 1.1), providing an *instrumental* justification for the transitivity axiom. Hence the "money pump" argument. Suppose Betty has the preferences just described. If she prefers a taco to a cup of yogurt, there must be a trade of the following type that she will agree to. Alf tells her that he will give her a taco in return for her cup of yogurt and some quantity of money (say one cent). Since she strictly prefers the taco to the yogurt, there must be some amount of money she will hand over to Alf (along with her yogurt) in exchange for a taco. So she makes the trade. Alf then proposes another trade: in return for another one cent and the taco, he will give her a pizza. Since she strictly prefers the pizza to the taco, there must be some amount of money she will hand over to Alf (along with her taco) in exchange for a pizza. So she makes the trade. So at this point she has traded her cup of yogurt, her taco, and two cents for a pizza. This makes sense, since she prefers pizza. But now Alf makes another offer: in exchange for her pizza and one more cent *he will give her yogurt back*. Since her preferences are not transitive, and so $z \succ x$ (she strictly prefers yogurt to pizza), she will make the trade. Now she is back where she began, with the yogurt, but she has spent three cents—all to get back to her original yogurt. And of course if Alf again offers to trade her a taco for her yogurt plus one more cent, again she will take the trade, and around and around she will go, serving as a money pump, making Alf richer and richer while she ends up where she started. So, it is said, we can see an instrumental or pragmatic justification for the transitivity axiom: agents who reject it could not possibly achieve their goals.

The money pump argument depends on the "more is better than less" axiom of *Homo Economicus* (Section 1.2): more money is better than less. That Betty ends up with less money is, other things equal, a bad outcome. Putting aside any worries that "more is better than less" may not hold for goods without qualification, the main worry is that the "more is better than less" axiom is itself an application of transitivity to amounts of goods. If quantity $2q$ of a good is better than quantity q, and if quantity $3q$ is better than $2q$, "more is better than less" requires that $3q$ is better than q. This, though, is just transitivity applied to quantities of goods. If one *really* questioned transitivity, one would also question whether more is better than less; and if so, then one would not be convinced by the money pump argument. If Betty holds that $q\$ \succ 3q\$$ the money pump argument won't move her.

This is not to deny that there is something deeply irrational about Betty; agents like her probably would have died out a long time ago. The money pump argument is persuasive in demonstrating to *us* how important transitivity is, but we should not expect it to move Betty. What it really shows is that *we* are deeply committed to transitivity.[13] It does not, though, provide an instrumental justification for transitivity if by that we mean a route to *accepting* transitivity, because only someone already committed to transitivity has access to the instrumental justification.

Rather than trying to provide instrumental or pragmatic justifications for the axioms of ordinal utility, it is better, I think, to see them as constitutive of our conception of a fully rational agent. Failure to recognize relations of transitivity is characteristic of schizophrenics;[14] those disposed to blatantly ignore transitivity are unintelligible to us: we can't understand their pattern of actions as sensible. ?? This is even more obviously the case with the *asymmetry* of strict preference. If someone prefers a pizza to a taco *and* a taco to a pizza, we just do not know what to make of his choices. To say that he would fail to satisfy his preferences, or be unsuccessful in practice, misses the point: we can't even understand what his preferences are. We cannot even make sense of ascribing a preference to an agent who does not conform to the *asymmetry* of strict preference.[15]

Some claim that the axioms of ordinal utility are more demanding than our understanding of a practically rational agent. Completeness seems especially strong and controversial. Completeness requires that for every possible pair of outcomes (x,y), $x \succeq y \geq y \succeq x$. But suppose the agent never has to choose between x and y; perhaps x and y only occur in the presence of z, and the agent always prefers z to both x and y. Say that x is a pizza with pepperoni, y a pizza with salami, and z a plain cheese pizza; perhaps our plain-cheese-loving pizza eater just has no preference relation between pepperoni and salami pizzas, but this doesn't matter, since she never has to make a choice between them. If we are impressed by such cases we may insist that a rational agent simply be able to have a *choice function* over *options* such that for any set of options (x,y,z), the agent can select the *best* option—that which is preferred to all others (see further Section 5.3). It looks as if our plain-cheese-loving pizza eater has such a choice function even though for her $\neg (x \succeq y \veebar y \succeq x)$. But unless the agent is able to relate all options, even her ability to choose may break down. If she goes to the Philosophy Department's Christmas party and finds only pepperoni and salami pizzas she will not be able to choose. Because she does not have a complete ordering she cannot say that pepperoni is worse than salami, better than salami, or even that she is indifferent between them. She just cannot relate them at all. For her, the choice between pepperoni and salami pizza is *incommensurable*: should she be confronted with those options she simply has no way to choose between them.[16] It is this that makes her look potentially irrational as a chooser. If we require that a person *always* has a choice function open to her (over all possible sets of options there is always a best choice), then she must conform to completeness.[17]

2.3 CARDINAL UTILITY THEORY

The Axioms

We have seen that an ordinal utility function for a person can be generated if her rankings satisfy completeness, asymmetry of strict preference/symmetry of indifference, reflexivity, and transitivity. But recall Table 2-1: ordinal utility function A, which numerically represents the options $(x, y, z,)$ as (3, 2, 1) is equivalent to ordinal utility function C, which represents them as (1000, 999, 1). We cannot say whether option y is "closer" to x or z: the numbers only represent the ordering of the options. We can get some idea of the *relative preference distances* between the options (roughly, *how much* one thing is preferred to another) by developing *cardinal utilities*, using some version (there are several) of additional axioms. On one accessible view, four further axioms are required. The key to this approach—pioneered by John von Neumann and Oskar Morgenstern—is to assume certain preferences over lotteries (risky outcomes), and then confront an agent with lotteries involving her ordinal outcomes.[18] Her ordinal preferences *over the lotteries* allow us to infer a cardinal scale (or, rather, as we shall see, a set of such scales). This is an incredibly powerful idea: it generates a cardinal utility measure from a series of ordinal preferences.[19]

One version of the axioms goes like this. In addition to the four axioms of ordinal utility we have just examined, we also need:

5. *Continuity*. Alf's preferences must be *continuous*. Suppose Alf has ranked three possibilities: having a pizza, having a taco, and having a cup of yogurt. Now suppose we give Alf a taco (his middle choice). He has the taco, but now we offer him a gamble: he can give up his taco and take a lottery ticket, in which the good prize is his first choice and the booby prize is his third choice (a cup of yogurt). Now we can easily imagine him rejecting many possible lotteries and keeping his taco. For example, suppose I offer him a lottery that gives him a .01 chance of getting a pizza and a .99 chance of getting a cup of yogurt. He probably will say, "thanks, but no thanks; I'll keep my taco." But suppose I offer him the opposite: a lottery that gives him a .99 chance of getting the pizza and only a .01 chance of getting the yogurt. Now we wouldn't be surprised if he gave up his taco for the lottery ticket: after all, he does prefer a pizza to a taco. For Alf's preferences to be continuous, it has to be the case that there is always some lottery in which the chances of getting his first choice and ending up with his third choice are such that he is indifferent between keeping his taco and accepting the lottery ticket. A little more formally, we can say that for all options (x,y,z) where $x \succeq y$ & $y \succeq z$ there must exist some lottery L that gives Alf a probability p of getting x (and so a $1\text{-}p$ of getting z) such that he is indifferent between having y and playing L.

6. *Better prizes.* Imagine that Alf is now confronted with two lotteries. In each lottery he is certain to end up with one of two prizes. The first lottery, say, is between a pizza and a cup of yogurt. The second lottery is between a taco and a cup of yogurt. Suppose the lotteries have the same probabilities of prizes: in Lottery 1 there is a .6 chance of a pizza and a .4 chance of a cup of the yogurt; in Lottery 2 there is a .6 chance of a taco a .4 chance of the yogurt. To conform to better prizes, Alf must prefer Lottery 1: when we compare the lotteries we see that they offer equal chances of winning the good prize (.6) and they offer equal chances of ending up with the bad prize (.4). Now in these lotteries the bad prize is the same, but in Lottery 1 the first prize is better, since Alf prefers a pizza to a taco. According to this axiom, Alf prefers to play the first lottery. Let us say (again, a little more formally) that if (i) Alf is confronted with lotteries L_1 over (w,x) and L_2 over (y,z); (ii) L_1 and L_2 have the same probability of prizes; (iii) the lotteries each have an equal prize in one position; (iv) they have unequal prizes in the other position; then (v) if L_1 is the lottery with the better prize, then for Alf $L_1 \succ L_2$; if neither lottery has a better prize, then for Alf $L_1 \sim L_2$.[20]
7. *Better chances.* Imagine that Alf is again confronted with two lotteries. In each lottery he is certain to end up with one of two prizes. Both lotteries are between a pizza and a cup of yogurt. In Lottery 1 there is a .7 chance of a pizza and a .3 chance of a cup of the yogurt; in Lottery 2 there is a .6 chance of a pizza and a .4 chance of the yogurt. To conform to better chances, Alf must prefer Lottery 1: the prizes are the same, but Lottery 1 gives him a better chance of his more preferred prize. So (i) if Alf is confronted with a choice between L_1 and L_2, and they have the same prizes, (ii) if L_1 has a better chance of the better prize, then for Alf $L_1 \succ L_2$.
8. *Reduction of compound lotteries.* If the prize of a lottery is another lottery, this can always be reduced to a simple lottery between prizes. This eliminates utility from the thrill of gambling: the only ultimate concern is the prizes.

If Alf meets these conditions, we can convert his ordinal utilities into cardinal utilities, which not only give the ordering of the payoffs but the size of the differences in the payoffs for each (or, more strictly, the ratios of the differences) where the higher the number, the better the outcome.

To grasp the crux of this method of generating cardinal utilities, assume that we have our three options: a pizza (x), a taco (y), and a cup of yogurt (z), where $x \succ y \succ z$ and we define the best option (x) as having a utility of 1, and the worst, (z), as 0. The question, then, is where on the scale of $1-0$ we should place y, the taco. If we were dealing simply with ordinal utilities, any number less than 1 and greater than 0 would suffice: but the idea is to get some notion of the amount of "preference distance" between, on the one hand, the taco, and on the other, the pizza and the cup of yogurt. Suppose that Alf is confronted with a lottery which gives him a p chance of getting the pizza and a $1-p$ chance of getting the yogurt. If he wins, he gets his pizza and if he loses he gets the cup of yogurt. Now we give him a binary choice: he can either have y, the taco (for certain), or he can play the lottery. It seems that Alf is very likely to prefer playing the lottery, when it gives a near 1 (perfect) chance of getting the pizza and a minute chance of getting the yogurt, to the certainty of the taco. In that case, he is essentially trading his second choice for the near certainty of his first choice. As p (the probability of winning the lottery) decreases toward zero, we would expect Alf to prefer to keep his taco (the certainty of getting his second choice) to a lottery that gives a tiny chance of a pizza and a very large chance of the booby prize—the cup of yogurt. At some point in between, as I have said, the continuity axiom says there is a value of p for which Alf is indifferent between the lottery $[L(x,z)]$ and y.

Suppose it turns out that he is indifferent between keeping y (his second choice) and playing a lottery that gives him a p of .9 of getting x and .1 chance of getting z. What we infer from this is that it takes a very large chance of getting his first option (.9) to induce Alf give up his second. He must, then, see y (the taco) as pretty good, if he will only play the lottery when he has a very great probability of winning. So we can say that on our scale of 1 (x, the pizza) to 0 (z, the yogurt), y, the taco, is at .9. In contrast, suppose that Alf was indifferent between having the taco for certain and playing a lottery than gave him a small chance (say .1) of getting the pizza and a .9 chance of ending up with the yogurt. From this we can infer that the taco must not be much better than the cup of yogurt, but the pizza must be a lot better: so we now give the taco a score of .1. We thus can generate a measurement in which the ratios between the numbers are significant from purely binary (ordinal) preferences involving lotteries.

I have said that the new cardinal measures tell us something about the "preference distance" between the options, but this interpretation is resisted by some. If we wish to be *extremely* careful, we will restrict ourselves to saying that all these "von Neumann–Morgenstern" utilities tell us are a person's preferences between lotteries or gambles, and so what he will do in certain situations that involve *risk*. That is, situations in which the chooser does not know for certain what outcome-consequences are associated with his action-options, but can assign a specific probability p that a certain action-option α will produce a certain consequence C_1.[21]

Questioning the Axioms

The von Neumann–Morgenstern axioms are especially controversial: there are well-known paradoxes associated with them and they are the object of continued debate. Consider first a simple objection. According to the continuity axiom there always must be some lottery L in which a rational agent is indifferent between certainty of keeping y and playing L, which has x and z as prizes. As R. Duncan Luce and Howard Raiffa acknowledged in their classic book on decision theory, some choices may not be continuous. To use their example: even if we all agree that $\$1 \succ 1¢ \succ$ death, not too many people are indifferent between 1¢ and a lottery with chance p of \$1 and a $1-p$ chance of death.[22]

A more complex objection, in this case to the better prizes axiom, is discussed by James Drier:

> Suppose you have a kitten, which you plan to give away to either Talia or Horace. Taila and Horace both want the kitten very much. Both are deserving, and both would care for the kitten. You are sure that giving the kitten to Taila [x] is at least as good as giving it to Horace [y, so $x \succeq y$]. But you think that would be unfair to Horace. You decide to flip a fair coin: if the coin lands heads, you will give the kitten to Horace, and if it lands tails, you will give the kitten to Talia.[23]

The problem is that you seem to have violated the better prizes axiom, according to which, it will be recalled, if (i) you are confronted with lotteries L_1 and L_2; (ii) L_1 and L_2 have the same probability of prizes; (iii) the lotteries each have an equal prize in one position; (iv) they have unequal prizes in the other position; then

(v) if L_1 is the lottery with the better prize, then $L_1 \succ L_2$ (in the story, $x \succeq y$.) To see the problem, suppose that L_1 has the prizes (x,z) and L_2 has the prizes (y,z), where z is simply a variable for the same outcome. Suppose further that L_1 and L_2 both give a .5 probability of winning z, and so there must be a .5 probability of winning the other prize (either x or y). L_1 and L_2 have equal prizes in the second position, so one's concern is just the first position. Since $x \succeq y$ (it is at least as good to give the kitten to Taila as to Horace), then according the better prizes axiom, $L_1 \succeq L_2$. Now let us substitute for the variable z a particular prize: x (Talia gets the kitten). So now L_1 is a .5 chance of (x,x) [that is, x—that Taila gets the kitten—for certain, since it is the prize in both positions] and L_2 a .5 chance of (y,x) [that is, a .5 chance that Horace will get the kitten and a .5 chance that Talia will]. In the first lottery (heads it's Talia's kitten, tails it's Taila's kitten); in the second lottery (heads it's Horace's kitten, tails it's Talia's). By better prizes, one prefers the first lottery. But this violates one's commitment to justice through a fair lottery; the person concerned with fairness holds that $L_2 \succeq L_1$, so better prizes is violated.

We again confront the deep issue of how to identify the correct description of the outcomes and options (see Section 2.4). Still, I think, however we characterize the outcomes, it looks like a rational person should conform to better prizes in this case. Suppose first that the only relevant differences between the outcomes concern who gets the kitten: all preferences are "who-gets-the-kitten" preferences. Now it looks as if the chooser ought not to violate better prizes by employing the fair lottery. To use the fair lottery to give away the kitten seems irrational if we suppose that *all you care about is who gets the kitten.* Why would you select a mechanism that sometimes gives the kitten to your preferred person and sometimes to the other *if the only thing you had preferences over was who ended up with the kitten?* So here violating better prizes seems objectionable. Assume, though, that you do not simply have preferences over "who-gets-the-kitten" but over "the process by which kittens are distributed." Here you opt for the fair lottery which can distribute to either Talia or Horace. Now the options may be better described as [*a*] "giving the kitten to the person who would be a better owner" and [*b*] "giving the kitten in a fair way," and you might hold that $b \succ a$.[24] If we understand the options in that way—that one of the things you have preferences over is the fairness of the process of distribution—the outcomes, and so the value of the action-options (your preferences over them), change and

TABLE 2-2 The Allais Paradox

		Payoffs		
	Options	Red Ball (1)	White Ball (89)	Blue Ball (10)
Lottery 1	A	1 million dollars	1 million dollars	1 million dollars
	B	zero	1 million dollars	5 million dollars
Lottery 2	C	1 million dollars	zero	1 million dollars
	D	zero	zero	5 million dollars

there is no violation of better prizes. So here, though it is rational to employ the fair lottery, employing it is consistent with better prizes.

The most famous challenge to the axioms of cardinal utility theory was presented by Maurice Allais.[25] Suppose that one is to draw a ball from an urn that has one red ball, eighty-nine white balls, and ten blue ones. Table 2-2 gives two pairs of lotteries.

Intuitively, we can see that according to better prizes and better chances, one's preferences over lotteries are to be determined only by differences in the size of the prize and the chance of getting it; if two lotteries have the same prize configurations and the same chances of winning the prize, then one will have the same preferences in the lotteries. Now in Lottery 1, your preference for option A could not be determined by the white ball, since both options give you the same chance of getting the same prize (an 89% chance of getting one million dollars). Better prizes and better chances tell us, when choosing between lotteries, to ignore in each the equal prizes with equal chances, and make our choice on the basis of better prizes and better chances. So if you do choose option A, then it must be the case that, in your estimate, the 10% chance of gaining an extra four million dollars in option B should the blue ball come up does not make up for the 1% chance of getting one million less in option B if the red ball comes up. So (roughly) if you choose A, you essentially prefer a gamble that, out of every eleven times, you get one million each time to a gamble that, out of every eleven times, you get five million ten times and nothing once.

If this is your reasoning, then you must also prefer option C in Lottery 2. Again, your choice cannot be made on the basis of what happens if the white ball comes up, since there are equal prizes with equal chances in both lotteries. Everything turns on the prizes and chances if the red or blue balls come up, but these are exactly the same prizes and chances as they are in Lottery 1. So the axioms commit you to option C. But many people who take option A in Lottery 1 take option D in Lottery 2. In Lottery 2, the idea of getting five million dollars ten out of eleven times and nothing one in eleven times seems like a reasonable bet, but it doesn't seem like a reasonable bet in Lottery A. And that seems to be because in Lottery 1, if one chooses A one is certain of getting a million dollars *no matter what happens*, and people have a hard time turning down the certainty of a million dollars. In contrast, in Lottery 2, there is no certain outcome and one is forced to gamble, and then people do seem to prefer a good chance of getting five million dollars, at the cost of a small chance of getting nothing.

This, though, means that what makes people choose differently in Lotteries 1 and 2 are the prizes concerning the white balls, but we have seen that since in both lotteries the white ball has equal chances of equal prizes, it should not affect one's choice between A and B, or between C and D. The issue, then, is whether people's tendencies to select A and D show that the axioms of cardinal utility are flawed insofar as rational people make choices that violate them, or whether we are often irrational in the way we judge probabilities. A crucial question here is whether rational people only seek to determine how well they might do, or whether rational people also seek to avoid regret.[26] In Lottery 1, if we select B and lose we might have deep regrets—"I had a million for sure and now I have nothing!"; but in Lottery 2 everything is a matter of chance, so we have little cause to regret our choice (we made a good bet and just had bad luck). As I see it, from the Allais Paradox we should not conclude either (a) that the axioms of cardinal utility fail to adequately capture our understanding of rational choice or (b) that those who choose A in Lottery 1 and D in Lottery 2 are irrational. Rather, it looks like people's utility functions—their rankings over outcomes—are often far more complicated than the monetary bets would indicate: one lottery leaves a person with the possibility of regrets and the other does not. As I will argue later in this chapter (Section 2.4), one's utility function can depend on the menu of options one faced, not just on the option one chose.

Clearly there are good questions that can be raised about cardinal utility theory. We can safely conclude that its axioms are far more controversial than the ordinal axioms: it is by no means hard to imagine rational agents who have noncontinuous preferences or who simply prefer to gamble (and therefore violate the reduction of compound lotteries). In general, however, I think philosophers have been rather too skeptical of the axioms: while some are rationally rejectable, they are not implausibly strong. The much-attacked better prizes and better chances axioms are not as vulnerable as is often thought. We also must not lose sight of the fact that the axioms are ways to generate cardinal measures out of ordinal preferences: ordinal *preferences* (meeting the axioms, of course) over outcomes and lotteries are all that are required. This is an especially elegant idea, but the very idea of cardinal utility does not depend on it. We should expect that doing something as neat as deriving the cardinal from the ordinal may invoke some contestable axioms. Although it is sometimes claimed that all uses of cardinal utility measures implicitly rely on the existence of the von Neumann–Morgenstern axioms,[27] in practice economists and game theorists are quite happy to appeal directly to the idea of a cardinal scale on which outcomes can be placed. Indeed, John Pollock has recently developed a computationally realistic model of rational decision making according to which cardinal, not ordinal, utility is fundamental. Pollock argues that a cognitive system that stored its basic values in an ordinal ranking would have to relate so many possible pairs of options that it would be unable to function—essentially such an agent would require an infinite data structure. (Pollock shows that a person who used pairwise comparisons to relate every possible state of the world over which she might choose would have more comparisons than the number of elementary particles in the universe!) Thus, rather than taking ordinal data as basic and trying to show how we might derive cardinal data from it, Pollock argues that real agents store their utility information in cardinal form.[28]

Pollock

Why Cardinal Measures Are Enticing

Why, a reader might ask, if there is so much dispute about cardinal utility functions, don't we content ourselves with ordinal utility? One reason, of course, is that some of us are not as skeptical about cardinal utility as are others: I have suggested that the criticisms are not as serious challenges as they are often thought to be. But another way to answer the question is to show why—if it could be achieved—cardinal utility is such an appealing idea. That would show us why, at least in the eyes of many, the worries and objections are not enough to make us run back to pure ordinalism! Suppose, then, that we did develop a cardinal measure of a person's utility. What could we do with it that we could not do with simple ordinal utility?

One of the problems we saw with simple ordinal utility was that we could not sensibly add the utility of different people into an overall, aggregate measure of utility. Think back again to Table 2-1: given the very different sets of numbers that represent the same preference structure, it was clear that ordinal utility functions do not lend themselves to addition. To add we need a cardinal measure. Now suppose that in Alf's utility function option *y* gives him .7 utility and *y* gives Betty .5: can we now proceed to add these nice cardinal numbers together, and say that the total utility of *y* is 1.2? Can we say that Alf gets more utility than Betty from *y*? Not without a *lot* more argument. We have assumed arbitrary highs (1) and lows (0) for each person: there is nothing to say that Alf's score of 1 for his best option identifies the "same utility" as Betty gets from her best option, to which she gives a score of 1. Given that the end points cannot be equated as the same, none of the ratios of distances that we identify in between can be automatically identified. More formally, cardinal utility functions derived through our axioms are only unique up to a linear transformation. If our function is U then any function U', where $U' = aU+b$ (where a is a positive real number and b is any real number), gives exactly the same information about ratios of differences between the options, and so serves equally well to describe a person's preferences.[29] Because of this, summing the utilities identified by one of the functions is not meaningful without an independent account providing a rationale of how they should be combined and at what ratios. There is no reason to suppose that Alf's .5 = Betty's .5; one might have an interpersonal measure that equates Alf's .5 with Betty's .75. Ken Binmore insists that "the problem isn't at all that making interpersonal comparisons is impossible. On the contrary, there are an infinite number of ways this can be done."[30] This is too strong, for there may be an infinite number of mathematical formulas for doing it but yet none might be justified. Certainly, though, the mere derivation of cardinal utility functions for each person does not tell us whether there is such a plausible function.

interpersonal compar

So, while cardinal utility might be inviting because some wish to add and compare different people's cardinal utilities, that looks more like a temptation to be avoided than a reason to embrace cardinal utility. What is genuinely inviting about cardinal utility is that it can be employed to perform expected utility calculations. Cardinal utilities have the *expected utility property*. Let us assume that Betty has a cardinal utility scale according to which the following outcomes are scaled: $w = 9$, $x = 8$, $y = 5$, $z = 3$. Suppose further that she is confronted with two action-options (α,β). Option α has two possible consequences (x,y); β has two possible consequences (w,z). We also need to suppose that Betty can assign probabilities to each outcome that would result from her performing the relevant act. Say that the probability of α producing x is .7; so the probability of α producing y must be $(1-.7)$, or .3 (since there is a probability of 1 that if she performs the act either x or y will occur, the probabilities must always sum to 1); similarly, if we assume that the probability of β producing outcome w is .5, the probability of producing z must also be .5. We can now calculate the expected utility of α and β using the formula that the expected utility ($E\mu$) of an action-option is the expected utility of its outcome multiplied by the probability that the outcome will be produced. Hence $E\mu(\alpha) = .7(8) + .3(5) = 7.1$; $E\mu(\beta) = .5(9) + .5(3) = 6$. Thus because $E\mu(\alpha) > E\mu(\beta)$, then $\alpha \succ \beta$. Based on her cardinal preferences over outcomes, Betty has been able to generate a preference over action-options even in cases where she is not certain what outcomes will be produced by her action-options. Notice that we can only make sense of expected utility theory by distinguishing a person's preferences over outcomes from her preferences over action-options (Section 2.1).

2.4 IS UTILITY THEORY A FORMALIZATION OF INSTRUMENTAL RATIONALITY?

No, It Isn't

Most see decision theory as an account of instrumental or a goal-oriented reasoning. Those who believe that all reasons are instrumental typically embrace decision theory because they think it is essentially a formalization of their view. Just as an instrumentally rational agent aims to maximize the satisfaction of her goals, it is thought, an agent who corresponds to the axioms of ordinal and cardinal utility theory seeks to maximize the satisfaction of her preferences. And if "goals" and "preferences" are the same thing, decision theory is simply a formal version of instrumental rationality. To be sure, the axioms add constraints on the structure of the preferences, but the core of the model is still seen as instrumental rationality. This, I think, is a serious mistake, albeit a common one.[31] Decision theory allows us to model choice based on one's notion of the overall ordering of outcomes by *whatever criteria one thinks appropriate*. What is required to generate a utility function is that one has some way to determine what is the best outcome, what is the next best outcome, and so on—but "best" need not be that which leads to the highest satisfaction of one's goals. There is no reason whatsoever to suppose that Alf's set of evaluative criteria are all about Alf's *goals*, *welfare*, or *goods* that he wishes to pursue.[32] Although decision theory distinguishes acts from outcomes (or consequences), and holds that the ranking of acts is determined by the ranking of outcomes, we should not confuse this sort of decision-theoretic consequentialism implicit in Figure 2-1 with the theory of instrumental action.[33] As Peter Hammond stresses, anything of normative relevance for choice is part of the consequence domain.[34] One of Alf's preferences over outcomes may be that he performs, rather than omits, act α, say "telling the truth when under oath today." If in his current set of options, one action-option is to tell the truth under oath, he will rank that act more highly than failing to tell the truth. Given this, the action of telling the truth under oath has "high utility"—that is, performing that action will "maximize his utility."[35] If, then, one's ranks outcomes on the basis of moral principles, a person acting on her moral principles can be modeled as maximizing a mathematical cardinal function.[36]

To better see how utility theory and instrumental rationality are distinct, consider the "ultimatum game." In this game, there is a good (say, an amount of money) to be divided between two players: in order for either player to get the money, both players have to agree to the division. In ultimatum games, the players make their moves sequentially. One player is selected by the experimenter to go first (call him the "Proposer"): the Proposer gives an ultimatum of the form: "I get x percent; you get y percent—take it or leave it!" No negotiation is allowed ($x+y$ must not exceed 100%). The second player is the "Disposer": she either accepts or rejects the offer. If she accepts, she gets her y percent, and the Proposer gets his x percent; if she rejects, neither gets anything. Now if we suppose that the players meet only once (they do not think they will ever play the game again), it would seem that the Proposer would propose 99% for himself, and 1% for the other. And it seems that Disposer, if she is instrumentally rational, would take the 1%. After all, as an instrumentally rational agent, she sees that 1% will achieve some of her goals, and 0% will not; so once the 99:1 offer has been made, as an instrumentally rational person it looks as if she must take it (more is, after all, better than less). And, as an instrumentally rational person, Proposer should see that it will better advance his ends to insist on 99%. But in experiments this does not happen: if Disposer is offered 1%, or 10%, or even 20%, it is very likely she will reject. And Proposer tends to demand "only" around 60% or so.[37] Does this mean that people act against their preferences, and so do not maximize their utility? I think not. The best explanation is that the players' utility functions are not simply about getting funds to best advance their goals, but about acting according to some norms of fair play. Gary E. Bolton has shown that, by building into a player's utility function (along with the goal of getting money) a concern for fairness to themselves (i.e., that the player is himself treated fairly), the actions of players in such games can be much better predicted.[38] But acting according to norms of fair play does not seem a goal: it is a principle to which a person wishes to conform.

It is true that some sorts of moral principled action cannot be modeled in terms of a cardinal utility function. One who is an "absolutist" about some principle, and so will never contemplate a lottery between acting on it and her second best option, violates continuity, and so we cannot develop a cardinal utility function for her. An absolutist still can have complete, reflexive, and transitive ordinal preferences (at least, so long as she has only one absolute principle).[39] The important point, though, *is that these sorts of worries cannot show that decision theory is about instrumental reasoning (or is instrumental in any interesting sense)*: they are objections to the lottery axioms and the development of *cardinal* utility. The difference between ordinal and cardinal utility regards the information implied about the relation between the ranked outcomes (not that cardinal utility commits us to instrumentalism but ordinal utility does not). Consequently, these problems with modeling some sorts of principled moral choices may be barriers to developing cardinal measures for the utility of such choices, but this by no means shows that moral choices based on principles or rules cannot be modeled in decision theory because it is inherently instrumental.

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

Should We Distinguish Preferences from Duty?

Amartya Sen dissents from my conception of decision theory: he advises us to distinguish actions that follow from "adhering to a deontological principle" from those that are "actually 'preferred.'"[40] The idea is that a moral obligation (say, to tell the truth) may require one to act in a way that sets back one's goals or welfare. Perhaps one's best friend will be convicted if one tells the truth under oath: his conviction is not an outcome "one prefers." Here Sen is pushing decision theory's notion of "preference" closer to its ordinary meaning of "liking" (see Section 2.1), where one can rationally do what one does not prefer ("I had reason to do it, but I sure did not prefer it.").[41]

Sen writes:

> A person's preferences over *comprehensive* outcomes (including the choice process) have to be distinguished from the conditional preferences over *culmination* outcomes *given* the act of choice. The responsibility associated with choice can sway our ranking of our narrowly-defined outcomes (such as commodity vectors), and choice functions and preference relations may be parametrically influenced by specific features of the *act* of choice (including the *identity* of the chooser, the *menu* over which the choice is made, and the relation of the particular *act* to behavioral social norms that constrain particular actions).[42]

Sen distinguishes the "comprehensive" outcome (which can include the utility of the choice process; for example, choosing in a fair way as in our case of Talia, Horace, and the kitten) from the distinct state of affairs that is produced by a choice, the "cumulative" outcome (who gets the kitten). Sen has in mind cases in which the utility of the states of affairs depends on the fact that one passed up what looked to be a more attractive option. Again, Sen:

> You arrive at a garden party, and can readily identify the most comfortable chair. You would be delighted if an imperious host were to assign you to that chair. However, if the matter is left to your own choice, you may refuse to risk it. You select a "less preferred" chair. Are you still a maximizer? Quite possibly you are, since your preference ranking for choice behavior may well be defined over "comprehensive outcomes," including choice processes (in particular, who does the choosing) as well as outcomes at culmination (the distribution of chairs).
>
> To take another example, you may prefer mangoes to apples, but refuse to pick the last mango from the fruit basket, and yet be very pleased if someone else were to "force" that last mango on you.[43]

Now on the face of it, this sort of chooser seems to act irrationally. Suppose one is confronted with the option {mango, apple}; given one's preference not to take the last mango, one will choose an apple. But now suppose that one is confronted with the set {mango, mango, apple}. Now one will pick a mango. This pattern of choices violates what many take to be two basic axioms of consistent rational choice—the contraction and weak expansion properties. According the *contraction* property, if *x* is chosen from the entire set *S*, it must be chosen from all subsets of *S* in which *x* is included. Our polite mango refuser violates this by selecting a mango from the set {mango, mango, apple} but an apple from the subset {mango, apple}.[44] Our chooser will also violate the *weak expansion* principle: if an option is chosen from each of two subsets, it must still be chosen when the sets are combined.[45] Suppose our person is confronted with two sets {apple, apple, mango} and {apple, mango}. Because she will not take the last mango, she will chose {apple} from the first set and {apple} from the second. But if we combine the two sets to get {apple, apple, apple, mango, mango} she will choose a mango, thus violating the weak expansion property.[46]

Supposing, as I think is clearly the case, that our "last-mango refuser" is not irrational, and so we want to allow for her preferences in an account of consistent choice, it may look as if we must follow

Sen in developing new axioms of rational choice. Sen seeks axioms that distinguish choices from menu-independent sets (where the contraction and weak expansion principles hold without modification) from choices involving options, like the choice of our mangoes, that are menu-dependent.[47] However, we need to recall our case of Talia, Horace, and the kitten (Section 2.3). Our polite last-mango refuser only violates the principles of consistent choice (contraction and weak expansion) if the choice is always viewed as over enjoyable food items. If Betty is simply picking the most enjoyable fruit, and if Betty chooses a mango when presented with the choice between a mango, an apple, and another apple, it is perplexing indeed if she then chooses an apple when confronted with the choice between a mango and an apple. It looks quite irrelevant that the first time her set included an extra apple (that she didn't want anyway). But, of course, the problem arises just because the relevant description changes (just as it did with our example of Talia, Horace, and the kitten): at one point Betty is choosing simply on the grounds of "Which fruit would I like the best?" and at the other time the relevant description is "Should I choose the one I like the best or be polite, knowing that Alf loves mangoes?" If Betty has reasons according to which, in cases like this, being polite is more important than an enjoyable fruit fest, then she is simply acting on her total set of preferences and there is no inconsistency.

The important point is that decision theory can model choices based on preferences over outcomes, where "preference" does not mean what one likes, but the outcomes that one has reason to choose to bring about. If one wishes to restrict "preference" to what one likes, or what promotes one's welfare, good, or goals, then we must follow Sen in distinguishing two preference orderings—those over "cumulative" and "comprehensive" outcomes. This in itself shows that the mere notion of a cardinal utility function says nothing about whether the maximization of one's utility is the same as the maximization of one's goals or aims (so again we see that utility theory is not a simply a version of instrumental rationality).

typo

The upshot is that, to formally model a purely instrumentally rational economic agent, we must not only embrace the axioms of formal decision theory that we have considered in the last two sections, but we must further constrain the agent's preferences so as to conform to the features of instrumentally rational agents and *Homo Economicus* that we examined in Chapter 1. Decision theory is a theory of rational choice; while decision theory can give us a formal utility function for *Homo Economicus*, it can also give us one for a principled moral agent.

Utility theory, then, is a much broader theory of rational agency than is *Homo Economicus*. The notion of economic rationality that we examined in the first chapter is based on instrumental rationality, more is better than less, decreasing marginal utility, downward sloping demand curves, etc. Though, I argued, it is more general than is often thought (selfishness, much less wealth maximization, is by no means a necessary trait of *Homo Economicus*), it is still a pretty specific conception of rational human action, which constrains the sorts of preferences a rational agent may have. Utility theory can model such preferences (so long as its basic axioms are met), but it can also model preferences that are based on principles of fairness, civility (not taking the last mango), and so on. Moral and political philosophers, then, should not confuse their (in my view justified) doubts that *Homo Economicus* is a general model of rational human action with (in my view unjustified) doubts that utility theory can be of use in their work.

2.5 DOUBTS FROM PSYCHOLOGY ABOUT EXPECTED UTILITY THEORY

Expected utility theory provides a highly formal and developed theory specifying how rational agents choose under conditions of *risk*—that is, where they are not certain about what consequences are produced by their action-options, but can assign probabilities relating each action-option and possible consequences. (If they cannot assign probabilities they are said to operate not under *risk*, but under *uncertainty*, which leads to yet further complications about rational expectations.) We have seen, though, that many people have reservations about the axioms, especially the better prizes and better chances axioms. People often seem to choose in ways inconsistent with their requirements. In the last twenty-five years cognitive psychologists, led by Daniel Kahneman and the late Amos Tversky, apparently have uncovered ways in which normal reasoners systematically violate the requirements of expected utility theory. In this section I briefly review some of their findings and then consider what implications they have for expected utility theory.

Biases and Heuristics

Errors in Probability Judgments The most basic and obvious problem is that most people are simply bad at making probability judgments: that is, even people of above-average intelligence do not rank outcomes in the way that expected utility theory would indicate. Consider:

> You are a fighter pilot who runs the risk of being killed by enemy fire. You can be killed in one of two ways: either by flak or by burns. You may also wear a jacket that will protect you entirely against one hazard, but is useless against the other, that is, you may wear a flak jacket or a burn jacket but not both. Two-thirds of the casualties result from flak; one-third from burns. You can wear either jacket all or part of the time. Which jacket do you choose to wear and why?[48]

Even pretty sophisticated reasoners who have taken courses in statistics tend to say "the flak jacket two-thirds of the time, and the burn jacket one-third of the time." But that will not maximize your chances of survival. Suppose there are 99 flights, each of which gets hit by enemy fire (we can ignore the flights that do not get hit). Assume all pilots wear the flak jacket two-thirds of the time: that is, for 66 missions. On those 66 missions, two-thirds of the deaths will be prevented (those from flak) while a third will die. So on those 66 missions, there will be 22 deaths. What about the remaining 33 missions (those for which only burn jackets are worn)? Here one-third will be saved (11) and two-thirds will die (22). So altogether, the two-thirds/one-third strategy will yield 22 + 22 deaths, or 44, which clearly is worse than wearing the flak jacket all the time, which will result in one-third of 99, or 33, deaths. But people have a strong tendency to respond to mixed threats with mixed responses, even though in cases like this a single response is best.

Even highly trained people make these sorts of errors, especially when they have to calculate probabilities given base rates in the population. Consider a simple problem posed by Richard Nisbett and Lee Ross:

> The present authors have a friend who is a professor. He likes to write poetry, is rather shy, and small in stature. Which of the following is his field (a) Chinese studies or (b) psychology?[49]

Tversky and Kahneman's research indicates that people will overwhelmingly select (a). The diagnostic information is *representative* of a

TABLE 2-3 Likelihood That One Has a Rare Disease after Testing Positive

	Have disease (10)	Don't have (9,990)
Test +	9.9	99.9
Test –	.1	9,890.1

professor of Chinese studies: people tend to be quite certain that the friend is a Chinese scholar. Yet, if we consider the relative size of the two populations—professors of psychology and professors of Chinese studies—the probability is very much that the person is a psychology professor. To be sure, the diagnostic information (i.e., the specific description of the friend) would justify some small departure from the probabilities given by the base rates, but the evidence indicates that in such situations people tend to wholly ignore base rate information, *even when it is supplied to them.*[50] Tversky and Kahneman conclude that "people's intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well."[51]

This bias can lead to serious errors when people rely solely on probability estimates of the accuracy of medical tests and ignore the base rates of the disease (or characteristic) in the population.[52] Suppose we have a relatively rare disease, say one that occurs at a rate of 1 in 1,000 (or, equivalently, 10 in 10,000). Suppose further that we have a test for the disease which is 99% accurate. We administer it to everyone in a population of 10,000. You test positive. Is it likely you have the disease? No, as Table 2-3 shows.

Of the entire randomly selected population who test positive, there is still only around a one-tenth chance that any one of them has the disease. Many people find this extremely surprising; if you do, then you will have trouble applying expected utility theory.

Another source of error in probability judgments is that "information is weighted in proportion to its vividness." Thus, for instance, concrete or emotionally salient information is more vivid, and hence is apt to play a dominant role in deliberating. Consider Nisbett's tale:

> Let us suppose that you wish to buy a new car and have decided that on the grounds of economy and longevity you want to purchase one of those solid, stalwart, middle-class

> Swedish cars—either a Volvo or a Saab. As a prudent and sensible buyer, you go to *Consumer Reports*, which informs you that the consensus of their experts is that the Volvo is mechanically superior, and the consensus of the readership is that the Volvo has a better repair record. Armed with this information, you decide to go and strike a bargain with the Volvo dealer before the week is out. In the interim, however, you go to a cocktail party where you announce this intention to an acquaintance. He reacts with disbelief and alarm. "A Volvo! You've got to be kidding. My brother-in-law had a Volvo. First, that fancy fuel injection computer thing went out. 250 bucks. Next he started having trouble with the rear end. Had to replace it. Then the transmission and the clutch. Finally sold it in three years for junk."[53]

Nisbett acknowledges that this gives you a reason to make a very small adjustment in the repair rates given by *Consumer Reports*; assuming that it wasn't in the original survey, you now have one additional observation. But is it likely to be weighed that lightly? More to the point, would you have the nerve to go out and buy a Volvo? This bit of information is so vivid that it is apt to drive out the bland statistics found in *Consumer Reports*.

Prospect Theory One of the von Neumann–Morgenstern axioms (Section 2.3) requires that people do not have preferences over whether to gamble, but only over outcomes. What has been dubbed "prospect theory" casts doubt on whether actual agents meet this condition. People show a marked tendency to accept risks about possible gains, but are much more averse to risk when it comes to possible losses. Consider the following gambles in Table 2-4.[54]

In both cases the expected utility is $5, but 55 of 132 subjects accepted one gamble and rejected the other. Of those that did so, 42 (out of the 55) rejected Gamble 1 but accepted Gamble 2. One difference seems to be that 1 invokes the possibility of loss, while 2 is about ways of gaining (something similar might be going on in the Allais Paradox in Table 2-2; given that people are sure to walk away with a million in option A, they may feel they might lose *their* money if the bet turns out badly in option B). People generally appear to put far more value on not losing x than on gaining x. If so, what gambles they take depend not just on the value of the prizes and the probabilities, but on whether the prize involves a loss or a gain.

TABLE 2-4 An Example of Prospect Theory

1. Would you accept a gamble that offers a 10% chance to win $95 and a 90% chance to lose $5?
2. Would you pay $5 to participate in a lottery that offers a 10% chance to win $100 and a 90% chance to win nothing?

This is especially striking in what is called the "endowment" effect. In one experiment students were given a free coffee mug, and asked whether they wanted to exchange it for a Swiss candy bar of roughly equal market value. About 10% of the students elected to give up the mug for the candy. In another group, the students were given the candy, and were offered the mug in exchange; again, about 10% of the students made the trade. Finally, in a third group no initial distribution was made, and students could choose either a mug or a candy bar; they split roughly equally in their choices.[55] This is striking, and poses a real worry about the whole idea of indifference curve analysis. Recall that indifference curves chart a person's preferences between bundles of goods; a person is indifferent between any bundles on the same curve. But the endowment effect suggests that one will prefer a mug to a candy bar if one now has the mug but switch to a preference for a candy bar over a mug if one presently has the candy bar. One could see this as a case of crossing indifference curves (which violate the fundamental condition of the asymmetry of strict preference) as Figure 2-2 shows.

If one starts at C (with 1 candy bar), one is only indifferent between it and some quantity of mugs greater than 1; if one starts with a mug, then one is indifferent between it and some quantity of candy bars greater than 1. One strictly prefers a mug to a candy bar and strictly prefers a candy bar to a mug! Such indifference curves are impossible given our understanding of rationality.

Framing Effects The example of the two identical bets in Table 2-4 is also an example of "framing effects": different ways of putting the "same choice" can yield different preferences over options. Consider Table 2-5, which shows another example (the percentages in parentheses are those who select this option).[56]

The pair A,C will result in the same number of lives saved and lost; the pair B, D will also result in the same number of lives saved and lost. A and C are just different descriptions of the same program, yet when

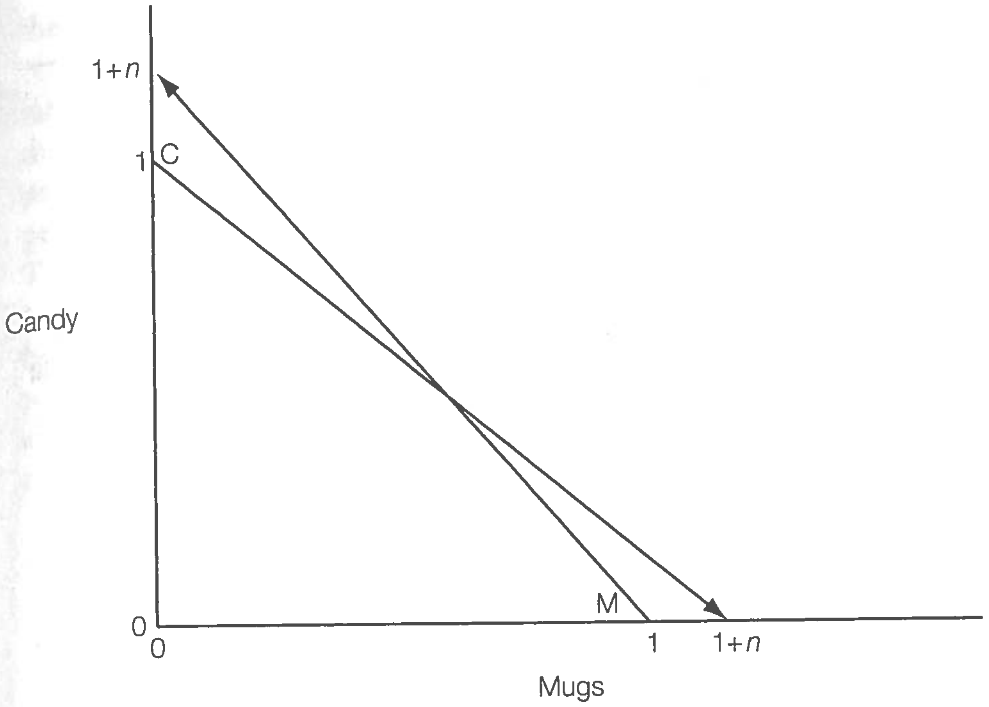

FIGURE 2-2 Crossing Indifference Curves and the Endowment Effect

the program is described in terms of saving lives, as it is in A, 72% of the respondents endorse it; when it is framed in terms of losing lives (as in C), only 22% endorse it. Similarly, although B and D are the same program, only 28% endorse B while 78% choose D. People are apt to make radically different choices depending on the way the choice is "framed" or described—saving lives or letting people die.

If one's choices are "framed" in this way—if different descriptions of the *same option* yield different utility[57]—the choices violate what Kenneth Arrow calls "extensionality":

> The cognitive psychologists refer to the "framing" of questions, the effect of the way they are formulated on the answers. A fundamental element of rationality, so elementary that we hardly notice it, is, in logicians' language, its *extensionality*. The chosen element depends on the opportunity set from which the choice is to be made, independently of how it is described.[58]

That is, the options must be stable in the sense that they describe *outcomes*, and people will have their preferences over action-options determined only by the outcomes associated with each option, not the way in which those outcomes are described.

TABLE 2-5 An Example of Framing Effects

1. Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

   A. If program A is adopted, 200 people will be saved. (72%)

   B. If program B is adopted, there is a one-third probability that 600 people will be saved and two-thirds probability that no people will be saved. (28%)

2. The same basic story is told with the following options:

   C. If program C is adopted, 400 people will die. (22%)

   D. If program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. (78%)

Do These Findings Undermine Expected Utility Theory?

The findings of cognitive psychologists such as Kahneman and Tversky must give pause to any advocate of expected utility theory: they point to well-documented shortcomings in people's ability to calculate probabilities and make choices based on them. However, to evaluate just how much of a challenge they pose, we need to distinguish several different ways in which they might lead us to doubt our account of rationality.

Certainly the often-replicated findings about the systematic errors people make in probability judgments show that expected utility is an idealization that most individuals never fully approach. That, though, should not be a great surprise: to assume rational choice is to assume a certain sort of ideal chooser, which perhaps few agents ever fully achieve. The question is whether the idealization is so far removed from reality as to be useless. If people are really *awful* at probability judgments, then it will not help a lot to try to understand their actions in terms of maximizing expected utility. It is not clear, though, that the findings are as troublesome as they first appear. As Richard Epstein points out, market competition provides a feedback mechanism that helps to correct erroneous judgments.[59] There is also evidence that, while people tend to be bad at calculating probabilities, they are much better at estimating frequencies and drawing the right conclusions about them.[60] Think again of our case of testing for the rare disease in Table 2-3. People seem to have a hard time thinking of the case in terms of probability calculations involving not only the probability that the test is right but also the probability that a random person in the population has the disease. But once put in terms of Table 2-3 the reasoning is clear. So too with the flak jacket example: a case that is puzzling to many when put in terms of probability becomes much easier when redescribed in terms of frequencies. This suggests that people may be considerably better at making the probabilistically correct choices when they are able to conceive of the choice in terms of frequencies.

Endowment effects pose more of a challenge for *Homo Economicus* than for expected utility theory. Economists typically (though not always) suppose that consumers simply have preferences over goods but not preferences whether they move from a certain starting point. *If* our preferences were only over goods, then endowment effects imply the deeply irrational indifference curves of Figure 2-2, where an individual prefers a candy bar to a mug *and* prefers a mug to a candy bar. But, of course, the crux of the issue is that the individuals do not simply have preferences over goods, but prefer to keep what they have to getting something else. Such preferences may be basic to what it means to "own" something; once you see something as your property, you may be reluctant to give it up, just because it is yours. "It ain't much, but it's mine" suggests that its being yours makes it more valuable. Having such preferences may be important to living a happy life; having them is apt to make each more pleased with the goods she ends up with, which she wouldn't trade "even for a lot of money." Again, to the extent that endowment effects are strong, economists may have to weaken their assumption that preferences are only over goods, but that is not a challenge to expected utility theory per se.

We are back to Talia, Horace, and the kitten (Section 2.3). If all preferences are over outcomes characterized independently of process, then there is something odd going on. But if agents have preferences not only over outcomes but also over the processes that produce the outcomes (Was the kitten given away by a fair lottery? Was the mug something of mine that I have to give up to get the candy?), then the oddness disappears. This, finally, leads us to the most fundamental issue: *framing*. Arrow, remember, argues for *extensionality*: preferences over outcomes must be independent of our description of them, and under framing we see that our evaluation of the "same" outcome changes as the description changes. Is this so? Think of our person who refuses to take the last mango. Can we say that she *really* has a choice between eating a mango and an apple, but she responds to different descriptions and so changes her preferences, and that is why her choices violate the contraction and weak expansion properties? I think it is clear that there is no such thing as a set of brute action-options that is independent of the descriptions (intentional states) of the choosers (Section 2.1). Are Betty's true options: a mango or an apple to eat, a soft object or a hard one, a dull-surfaced object or a shiny-surfaced one, the superior piece of fruit to throw at a disliked political speaker, the superior fruit to put on the teacher's desk, or between being rude and being civil? One of the hopes of revealed preference theory, with its behavioral underpinnings, was that we could describe an unambiguous "choice behavior" that had no reference to the chooser's intentional states, and so her descriptions of what she is up to. But as I have argued, this behaviorist project failed: action is inherently intentional. So "framing" cannot simply be understood in terms of different descriptions of the "same" option, for what is the "same" option depends on the relevant description.[61] On Sen's view framing explains *inconsistent choices* but, as he points out, our person who refuses to take the last mango does not really seem inconsistent.[62]

To better to see the complexities, suppose that when possible state of the world W (spatiotemporally defined) is described as D Betty gives it μ utility, but when W is redescribed as D' she gives it μ' (where $\mu' > \mu$), even when the truth of D is consistent with the truth of D'. Under description D she sees her action-option as α; under D' she sees her action-option β (where $\beta \succ \alpha$). We cannot conclude that she has been subject to framing, for D' may have alerted her to a relevant description that changes her evaluation of W and her understanding of the action which brings it about ("it isn't just about choosing fruit, it is also about civility"). To show Betty is in some way irrational we might show that she has manifestly relied on an *irrelevant* consideration in changing her preference, or that she chooses differently when she thinks about the good aspects of the option and when she thinks about its bad features. In this latter case we would expect that her preferences will be inconsistent (sometimes $x \succ y$, other times $y \succ x$) depending on what she is thinking about: when she thinks about how many lives will be saved she prefers x to y, but when she thinks about how many lives will be lost she prefers y to x.

A full account of framing, and its relation to a plausible version of Arrow's condition of "extensionality," must then involve a notion of *irrelevant* differences in description or a criterion of choice inconsistency.[63] In our obvious framing case in Table 2-5 it seems that there is something amiss because there is apparently no good reason for drawing a distinction between A and C, or B and D, and yet the respondents do. But if there is a good reason for drawing the distinction, no framing occurs. Think about the cases in Table 2-4 which involve both the endowment effect and "framing" (the case involving buying a lottery ticket and making a bet). Suppose that the respondent supported state lotteries because they were used to fund education: now he might buy a lottery ticket for an expected payoff of $5 (he should be so lucky!) while turning down a bet with a payoff of $5, and have perfectly rational preferences. What this suggests, then, is that we need some account of which distinctions are relevant and which are not or, as John Broome says, what justifies a preference.[64] Underlying or justifying a preference ordering must be a system of principles, goals, ends, or values, and it is this that can justify distinguishing outcomes in terms of their descriptions. If this is so, then preferences over states of affairs cannot be basic. There are an infinite number of descriptions of any one state of affairs.[65] Our conviction that something is amiss in obvious cases of framing shows that we do not think every change in description makes for a different outcome. But, then, which do and which don't? It looks as if the only way to justify making a distinction is to draw on some other evaluative criteria to justify our preferences.

This shows, I think, that utility theory is a way to formalize and model rational action, but is not itself a complete theory of rational action. To employ utility theory presupposes that we know which are the relevant, and which are the irrelevant, features for evaluating states of affairs. Unless we possess such criteria we cannot distinguish framing effects from redescribing the world in such a way that we call attention to an important feature. However, only a value and/or a moral theory can allow us to do that; utility theory does not imply any specific value or moral theory, but presupposes that an agent employs one and so can rank the outcomes. One of the things I hope this chapter has made clear is that in formal utility theory, "utility" is not a sort of value, but simply a representation of one's ordering of options based on one's underlying values, ends, and principles.

CHAPTER 2

SUMMARY

This chapter has explained the basics of utility theory, and I have presented my own views in regard to several controversial questions. In this chapter I have:

- *Distinguished the inherently relational idea of a "preference" from notions such as "tastes" or "likings" with which it is often confused.*
- *Distinguished preferences over outcomes from preferences over action-options.*
- *Explained and defended the axioms of ordinal utility theory, and explained what is meant by an "ordinal utility function."* To have an ordinal utility function a person must have a complete ordering of the feasible options, her strict preferences must be asymmetric, her relations of indifference must be symmetric, and her preferences must be reflexive and transitive. An ordinal utility function is a numerical representation of the ordering of the options.
- *Explained, and generally defended, the axioms of cardinal utility theory.* In order to have a cardinal utility function a person must have preferences not only over outcomes but also over lotteries. Her preferences must be continuous and satisfy the better prizes, better chances, and reduction of compound lotteries axioms. I considered several paradoxes associated with the cardinal utility axioms; I argued that these paradoxes usually are the result of ascribing too simple a utility function to the choosers, or a too-simple description of the choices they are confronting.
- *Argued that (1) utility theory does not maintain that the aim of our preferences is to achieve utility, and that (2) utility theory is not simply a formalization of instrumental rationality.* Point (1) is generally accepted; point (2) is more controversial. Utility theory is a broad theory that can model both instrumental and noninstrumental rational action. In defense of point (2) I examined the way that considerations of principle can be modeled into utility functions. We will return to this important matter in Chapter 4.
- *Explained that cardinal utility has the expected utility property.*
- *Examined some of the main findings of social and cognitive psychologists about the ways that people fall short of the predictions of expected utility theory.* It is my view that these findings generally show that

people are imperfectly rational, but they do not undermine the usefulness of utility theory as a way to model human actions. Some of the findings show that people have difficulty with some ways of thinking about probabilities. Others, such as the endowment effect, once again point to the importance of not assuming too simple a view of what people's preferences range over.

- *Emphasized the importance of "framing."* I have argued that we can only distinguish the framing effect from a relevant difference in the description of an outcome or action by appealing to a value theory, or a moral theory, that identifies the choice-relevant features of states of affairs and actions. Utility theory does not do this, and so it is best understood as a formalization of rational action that presupposes a value or moral theory.

NOTES

1. See R. Duncan Luce and Howard Raiffa, *Games and Decisions*, p. 21. They did acknowledge that this is a very rough interpretation.
2. Louis Kaplow and Steven Shavell, *Fairness versus Welfare*, p. 431.
3. Michael Walzer, "Political Action: The Problem of Dirty Hands," pp. 166–167.
4. See S. I. Benn and G. W. Mortimore, "Technical Models of Rational Choice." pp. 160–161. Amartya Sen has developed an account whereby a rational person may be said to choose her less preferred outcome. See his "Maximization and the Act of Choice."
5. Cass Sunstein writes:

   > If we think of a preference as something that lies behind a choice, what is it exactly? How can it be identified or described? Internal mental states are extraordinarily complex, and the constellation of motivations that lies behind a choice in one setting may be quite different from the constellation that produces choice in a different time and place. People's decisions are based on whims, second-order preferences, aspirations, judgments, drives of various kinds, and so forth, each potentially coming to the fore depending on the context.

   All this is too complicated, Sunstein believes: it leads to all the "difficulties that the 'revealed preference' idea was supposed to overcome." Sunstein, *Free Markets and Social Justice*, p. 16. The classic formulations of revealed preference theory are by Paul Samuelson. See, for example, his

"Consumption Theory in Terms of Revealed Preference," and "A Note of the Pure Theory of Consumer Behavior."

6. For a formal account along these lines, see Peter Hammond, "Consequentialist Foundations for Expected Utility."
7. To make things more complicated we need full descriptions such as "the world in which Betty is a tax lawyer at a large firm and eats good food five times a week."
8. I am following Christopher McMahon here: "what there is best reason for an agent to do is determined by the value (from the agent's point of view) of the outcomes correlated with the available actions." *Collective Rationality and Collective Reasoning*, p. 7.
9. Alas, this is an oversimplification. Relevant here is the difference between the utility theory as articulated by L. J. Savage and that of Richard Jeffrey. The view presented in the text sounds more like that of Savage: the utility of the action derives directly from the utility of the states of affairs with which it is correlated. For Jeffrey, the act chosen may itself affect the utility of the resulting state of affairs. It would take us too far afield to go into these matters, though I do intend, by using the general idea of a "correlation" between action and outcome, to allow for conditional probabilities. For a nice summary of the difference between these two views, see Brian Skyrms, *Evolution of the Social Contract*, pp. 47–48. For an example of how the act chosen may itself affect the utility of the resulting state of affairs, see the discussion of the Newcomb problem in Section 4.2.
10. Recall Hume's statement in Section 1.1 about the relation of ends to reason; the proposals we considered that aim to "clean up" preferences are relevant here.
11. What is sometimes called a "strong" ordering has only strict preference relations (no indifference). So for all pairs of options, $x \succ y \veebar y \veebar x$.
12. Jean E. Hampton in *The Authority of Reason* takes this type of challenge very seriously. Indeed, she takes seriously the challenge "Why should I worry about being rational?" See also David Schmidtz, *Rational Choice and Moral Agency*, Chapter 1.
13. Benn and Mortimore argue that rationality itself does not require transitivity.
14. Michael Argyle, *The Psychology of Interpersonal Behavior*, 3rd edition, p. 211.
15. Benn and Mortimore, "Technical Models of Rational Choice," p. 163.
16. I discuss incommensurability as incompleteness of preference orderings in my *Contemporary Theories of Liberalism*, Section 2.3.
17. Amartya Sen, *Collective Choice and Social Welfare*, Chapter 1.
18. See John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior.*

19. My aim here is to give an intuitive idea of the axioms. I am primarily drawing on James Dreier, "Decision Theory and Morality," who stresses subjective utility. But see also Hampton, *The Authority of Reason*, Chapter 7; Luce and Raiffa, *Games and Decisions*, pp. 23–31.

20. See Luce and Raiffa, *Games and Decisions*, p. 27.

21. See James D. Morrow, *Game Theory for Political Scientists*, p. 34.

22. Luce and Raiffa, *Games and Decisions*, p. 27.

23. Dreier, "Decision Theory and Morality," p. 173. For other discussions of this problem, see John Broome, "Rationality and the Sure-Thing Principle," p. 90; Peter A. Diamond, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility: Comment."

24. This interpretation is considered by both Dreier in "Decision Theory and Morality" and Broome in "Rationality and the Sure-Thing Principle."

25. For helpful discussions, see Daniel M. Hausman and Michael S. McPherson, *Economic Analysis and Moral Philosophy*, pp. 33–35; Broome, "Rationality and the Sure-Thing Principle."

26. See Broome, "Rationality and the Sure-Thing Principle."

27. William Riker, *Liberalism Against Populism*, p. 95.

28. See John Pollock, *Thinking About Acting*, Chapter 2.

29. Hausman and McPherson, *Economic Analysis and Moral Philosophy*, p. 32.

30. See Ken Binmore, *Natural Justice*, p. 121.

31. For an extremely insightful if contentious analysis, see Hampton, *The Authority of Reason*, Chapter 7. David Gauthier makes the error of conceiving of decision theory as instrumental in *Morals by Agreement*, Chapter 2. Morrow presents a typical interpretation: "Put simply, rational behavior means choosing the best means to gain a predetermined set of ends." *Game Theory for Political Scientists*, p. 17.

32. Cf. Morrow, *Game Theory for Political Scientists*, p. 17.

33. As Paul Anand recognizes. *Foundations of Rational Choice Under Risk*, p. 84n.

34. Hammond, "Consequentialist Foundations for Expected Utility," p. 26.

35. S. I. Benn has modeled deontological requirements in this way. See *A Theory of Freedom*, Chapter 3.

36. As John Rawls notes: "A utility function is simply a mathematical representation of households' or economic agents' preferences, assuming these preferences to satisfy certain conditions. From a purely formal point of view, there is nothing to prevent an agent who is a pluralistic intuitionist from having a utility function." *Political Liberalism*, p. 332n. I defend this idea in much more detail in "Reasonable Utility Functions and Playing the Fair Way."

37. See Skyrms, *Evolution of the Social Contract,* Chapter 2.

38. Gary E. Bolton, "A Comparative Model of Bargaining: Theory and Evidence." Bolton treats money and fairness as substitutable. Benn and Mortimore agree that deontic preferences can be cardinalized using the lottery axioms. "Technical Models of Rational Choice," pp. 185–186.

39. See Benn, *A Theory of Freedom*, Chapter 3. If one has two absolutist principles it would seem that one would violate either completeness or asymmetry of strict preference. As Rawls notes, a strict lexicographic preference ordering prevents formulating a cardinal utility function. *Political Liberalism*, p. 332n.

→ 40. Sen, "Maximization and the Act of Choice," p. 191.

41. See Benn and Mortimore, "Technical Models of Rational Choice," pp. 160–161. 71

42. Sen, "Maximization and the Act of Choice," p. 159.

43. Ibid., 161, footnote omitted.

44. See Anand, *Foundations of Rational Choice Under Risk*, pp. 56–58.

45. I consider these principles in more detail in the context of social choice in Section 5.3. I consider there in just what way this is a "weak" expansion principle.

46. See Sen, *Collective Choice and Social Welfare*; Dennis Mueller, *Public Choice III*, pp. 152–153.

47. Sen's argument is complex. He argues for a notion of maximization that is distinct from optimization, which itself has to drop consistency conditions. I cannot go into these matters here. See "Maximization and the Act of Choice," especially p. 184n.

48. This example is recounted by Richard Epstein in his *Skepticism and Freedom*, p. 229.

49. Richard E. Nisbett and Lee Ross, *Human Inference*, p. 25.

50. See Daniel Kahneman and Amos Tversky, "On the Psychology of Prediction."

51. Tversky and Kahneman, "Belief in the Law of Small Numbers," p. 25.

52. For an application to the problems in law, see Deborah Davis and William C. Follette, "Rethinking the Probative Value of Evidence: Base Rates, Intuitive Profiling, and the 'Postdiction' of Behavior."

53. Quoted in Nisbett and Ross, *Human Inference*, p. 15.

54. Daniel Kahneman and Amos Tversky, "Choices, Values and Frames," p. 15. As Kahneman and Tversky note, the "framing effect" is also at work here, something we shall presently consider.

55. Jack L. Knetsch, "Endowment Effect and Evidence on Nonreversible Indifference Curves," pp. 172–173.

56. Kahneman and Tversky, "Choices, Values and Frames," p. 5.

57. See Amos Tversky and Daniel Kahneman, "Rational Choice and the Framing of Decisions," p. 211.

58. Kenneth J. Arrow, "Risk Perception in Psychology and Economics," p. 6.

59. Epstein, *Skepticism and Freedom*, pp. 228–232.

60. See Steven Pinker, *How the Mind Works*, pp. 347–348.

61. Anand is sensitive to these issues. See his *Foundations of Rational Choice Under Risk*, pp. 93–94.

62. Sen, "Maximization and the Act of Choice," p. 168n.

63. Arrow himself refers to people being moved by "irrelevant" events in his "Risk Perception in Psychology and Economics," p. 7.

64. See Broome, "Rationality and the Sure-Thing Principle."

65. See Stuart Hampshire, *Morality and Conflict*, p. 106.