1
Revealed Preference

1.1 Rationality?

A rational number is the ratio of two whole numbers. The ancients thought that all numbers were rational, but Pythagoras's theorem shows that the length of the diagonal of a square of unit area is irrational. Tradition holds that the genius who actually made this discovery was drowned, lest he shake the Pythagorean faith in the ineffable nature of number. But nowadays everybody knows that there is nothing irrational about the square root of two, even though we still call it an irrational number.

There is similarly nothing irrational about a philosopher who isn't a rationalist. Rationalism in philosophy consists of arriving at substantive conclusions without appealing to any data. If you follow the scientific method, you are said to be an empiricist rather than a rationalist. But only creationists nowadays feel any urge to persecute scientists for being irrational.

What of rational decision theory? Here the controversy over what should count as rational is alive and kicking.

Bayesianism. Bayesianism is the doctrine that Bayesian decision theory is always rational. The doctrine entails, for example, that David Hume was wrong to argue that scientific induction can't be justified on rational grounds. Dennis Lindley (1988) is one of many scholars who are convinced that Bayesian inference has been shown to be the only coherent form of inference.

The orthodoxy promoted by Lindley and others has become increasingly claustrophobic in economics, but Gilboa and Schmeidler (2001) have shown that it is still possible to consider alternatives without suffering the metaphorical fate of the Pythagorean heretic who discovered the irrationality of $\sqrt{2}$. Encouraged by their success, I follow their example by asking three questions:

What is Bayesian decision theory?

When should we count Bayesian decision theory as rational?

What should we do when Bayesian decision theory isn't rational?

In answering the first question, I hope to distinguish Bayesian decision theory from Bayesianism. We can hold on to the virtues of the former without falling prey to the excesses of the latter.

In answering the second question, I shall note that Leonard (Jimmie) Savage—normally acknowledged as the creator of Bayesian decision theory—held the view that it is only rational to apply Bayesian decision theory in small worlds. But what is a small world?

The worlds of macroeconomics and high finance most certainly don't fall into this category. What should we do when we have to make decisions in such large worlds? I am writing this book because I want to join the club of those who think they have the beginnings of an answer to this third question.[1]

No formal definition of rationality will be offered. I don't believe in the kind of Platonic ideal that rationalist philosophers seem to have in mind when they appeal to Immanuel Kant's notion of Practical Reason. I think that rationality principles are invented rather than discovered. To insist on an a priori definition would be to make the Pythagorean mistake of prematurely closing our minds to possible future inventions. I therefore simply look for a minimal extension of orthodox decision theory to the case of large worlds without pretending that I have access to some metaphysical hotline to the nature of absolute truth.

1.2 Modeling a Decision Problem

When Pandora makes a decision, she chooses an action from those available. The result of her action will usually depend on the state of the world at the time she makes her decision. For example, if she chooses to step into the road, her future fate will depend on whether a car happens to be passing by.

We can capture such a scenario by modeling a decision problem as a function

$$D : A \times B \to C$$

in which A is the set of available actions, B is the set of possible states of the world, and C is the set of possible consequences. So if Pandora chooses action a when the world is in state b, the outcome will be $c = D(a, b)$. Figure 1.1 illustrates a simple case.

An *act* in such a setting is any function $\alpha : B \to C$. For example, if Pandora bets everything she owns on number 13 when playing roulette,

[1] For some overviews, see Hammond (1999), Kadane, Schervish, and Seidenfeld (1999), and Kelsey (1992).

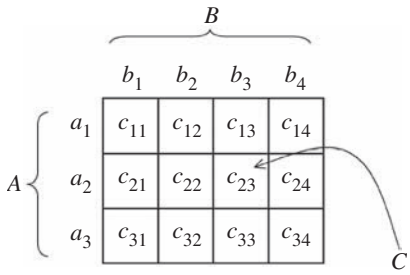

Figure 1.1. A decision problem. Pandora chooses one of the actions: a_1, a_2, a_3. Nature chooses one of the states: b_1, b_2, b_3, b_4. The result is a consequence $c_{ij} = D(a_i, b_j)$. The rows of the matrix of consequences therefore correspond to acts and the columns to states of the world.

then she chooses the act in which she will be wealthy if the little ball stops in the slot labeled 13, and ruined if it doesn't. Pandora's choice of an action always determines some kind of act, and so we can regard A as her set of feasible alternatives within the set $\aleph$ of all possible acts.[2]

If Pandora chooses action a from her feasible set A in a rational way, then we say that a is an optimal choice. The framework we have chosen therefore already excludes one of the most common forms of irrationality—that of choosing an optimal action without first considering what is feasible.

Knowledge. Ken Arrow (1971, p. 45) tells us that each state in B should be "a description of the world so complete that, if true and known, the consequences of every action would be known." But how does Pandora come to be so knowledgeable?

If we follow the philosophical tradition of treating knowledge as justified true belief, the answer is arguably never. I don't see that we are even entitled to assume that reality accords to some model that humans are able to envisage. However, we don't need a view on the metaphysical intelligibility of the universe to discuss decision theory intelligently. The models we use in trying to make sense of the world are merely human inventions. To say that Pandora knows what decision model she is facing can therefore be taken as meaning no more than that she is committed to proceeding as though her model were true (section 8.5).

1.3 Reason Is the Slave of the Passions

Thomas Hobbes characterized man in terms of his strength of body, his passions, his experience, and his reason. When Pandora is faced with a

[2] The Hebrew letter $\aleph$ is pronounced "aleph."

decision problem, we may identify her strength of body with the set A of all actions that she is physically able to choose. Her passions can be identified with her preferences over the set C of possible consequences, and her experience with her beliefs about the likelihood of the different possible states in the set B. In orthodox decision theory, her reason is identified with the manner in which she takes account of her preferences and beliefs in deciding what action to take.

The orthodox position therefore confines rationality to the determination of means rather than ends. To quote David Hume (1978): "Reason is, and ought only to be, the slave of the passions." As Hume extravagantly explained, he would be immune to accusations of irrationality even if he were to prefer the destruction of the entire universe to scratching his finger. Some philosophers hold to the contrary that rationality can tell you what you ought to like. Others maintain that rationality can tell you what you ought to choose without reference to your preferences. For example, Kant (1998) tells us that rationality demands that we honor his categorical imperative, whether or not we like the consequences.

My own view is that nothing is to be gained in the long run by inventing versions of rationality that allow their proponents to label brands of ethics or metaphysics other than their own as irrational. Instead of disputing whose ethical or metaphysical system should triumph, we are then reduced to disputing whose rationality principles should prevail. I prefer to emulate the logicians in seeking to take rationality out of the firing line by only adopting uncontroversial rationality principles.

Such a minimalist conception of rational decision theory isn't very glamorous, but then, neither is modern logic. However, as in logic, there is a reward for following the straight and narrow path of rectitude. By so doing, we will be able to avoid getting entangled in numerous thorny paradoxes that lie in wait on every side.

Consistency. The modern orthodoxy goes further than David Hume. It treats reason as the slave, not only of our passions, but also of our experience. Pandora's reason is assumed to be the slave of both her preferences and her beliefs.

It doesn't follow that rational decision theory imposes no constraints on our preferences or our beliefs. Everyone agrees that rational people won't fall prey to the equivalent of a logical contradiction. Their preferences and beliefs will therefore be consistent with each other in different situations. But what consistency criteria should we impose? We mustn't be casual about this question, because the words *rationality* and *consistency* are treated almost as synonyms in much modern work.

For example, Bayesianism focuses on how Pandora should respond to a new piece of data. It is said that Pandora must necessarily translate her prior beliefs into posterior beliefs using Bayes' rule if she is to act consistently. But there is seldom any serious discussion of *why* Pandora should be consistent in the sense required. However, this is a topic for a later chapter (section 7.5.2). We already have enough contentious issues in the current chapter to keep us busy for some time.

1.4 Lessons from Aesop

The fox in Aesop's fable was unable to reach some grapes and so decided that they must be sour. He thereby irrationally allowed his beliefs in domain B to be influenced by what actions are feasible in domain A.

If Aesop's fox were to decide that chickens must be available because they taste better than grapes, he would be guilty of the utopian mistake of allowing his assessment of what actions are available in domain A to be influenced by his preferences in domain C. The same kind of wishful thinking may lead him to judge that the grapes he can reach must be ripe because he likes ripe grapes better than sour grapes, or that he likes sour grapes better than ripe grapes because the only grapes that he can reach are probably sour. In both these cases, he fails to separate his beliefs in domain B from his preferences in domain C.

Aesop's principle. These observations motivate the following principle:

> Pandora's preferences, her beliefs, and her assessments of what is feasible should all be independent of each other.

For example, the kind of pessimism that might make Pandora predict that it is bound to rain now that she has lost her umbrella is irrational. Equally irrational is the kind of optimism that Voltaire was mocking when he said that if God didn't exist, it would be necessary to invent Him.

1.4.1 Intrinsic Preferences?

It is easy to propose objections to Aesop's principle. For example, Pandora's preferences between an umbrella and an ice cream might well alter if it comes on to rain. Shouldn't we therefore accept that preferences will sometimes be state-dependent?

It is true that *instrumental* preferences are usually state-dependent. One can tell when one is dealing with an instrumental preference, because it advances matters to ask Pandora *why* she holds the preference. She might say, for example, that she prefers an umbrella to an

ice cream because it looks like rain and she doesn't want to get wet. Or that she prefers driving nails home with a hammer rather than a screwdriver because it takes less time and trouble. More generally, any preference over actions is an instrumental preference.

One is dealing with *intrinsic* preferences when it no longer helps to ask Pandora why she likes one thing rather than another, because nothing relevant that might happen is capable of altering her position.[3] For example, we could ask Pandora why she likes wearing a skirt that is two inches shorter than the skirts she liked last year. She might reply that she likes being in the fashion. Why does she like being in the fashion? Because she doesn't like being laughed at for being behind the times. Why doesn't she like being laughed at? Because girls who are ridiculed are less attractive to boys. Why does she like being attractive to boys? One could reply that evolution made most women this way, but such an answer doesn't take us anywhere, because we don't plan to consider alternative environments in which evolution did something else. The fact that Pandora likes boys can therefore usefully be treated as an intrinsic preference. Her liking for miniskirts is instrumental because it changes with her environment.

Economists are talking about intrinsic preferences when they quote the slogan: *De gustibus, non est disputandum*. In welfare economics, it is particularly important that the preferences that we seek to satisfy should be intrinsic. It wouldn't help very much, for example, to introduce a reform that everyone favors if it changes the environment in a way that reverses everyone's preferences.

1.4.2 Constructing a Decision Problem

Decision problems aren't somehow built into the structure of the universe. Pandora must decide how to formulate her decision problem for herself. There will often be many formulations available, some of which satisfy the basic assumptions of whatever decision theory she plans to apply—and others which don't. She might have to work hard, for example, to find a formulation in which her preferences on the set C of consequences are intrinsic. If she plans to apply Aesop's principle completely, she must work even harder and construct a model in which there are no linkages at all between any of the sets A, B, and C (other than those built into the function D) that are relevant to her decision.

For example, if Pandora's actions are predictions of the weather, she mustn't take the states in B to be *correct* or *mistaken*, because the true

[3] The distinction between intrinsic preferences and instrumental preferences is made in economics by speaking of direct and indirect utility functions.

state of the world would then depend on her choice of action. In the case of an umbrella on a rainy day, it may be necessary to identify the set C of consequences with Pandora's *states of mind* rather than physical objects. One can then speak of the states of mind that accompany having an umbrella-on-a-sunny-day or having an umbrella-on-a-wet-day, rather than speaking just of an umbrella.

Critics complain that the use of such expedients makes the theory tautological, but what could be better when one's aim is to find an uncontroversial framework?

Every thing is what it is, and not something else. The price that has to be paid for an uncontroversial theory is that it can't be used to model everything that we might like to model. For example, we can't model the possibility that people might choose to change their intrinsic preferences by adopting behaviors that are likely to become habituated. Such restrictions are routinely ignored when appeals are made to rational decision theory, but to bowdlerize Bishop Butler: Every theory is good for what it is good for, and not for something else.

To ignore the wisdom of Bishop Butler is to indulge in precisely the kind of wishful thinking disbarred by Aesop's principle. We aren't even entitled to take for granted that Pandora *can* formulate her decision problem so that Aesop's principle applies. In fact, I shall be arguing later that it is a characteristic of decision making in large worlds that agents are *unable* to separate their preferences from their beliefs. However, this unwelcome consideration will be left on ice until chapter 9.

Finally, nothing says that one can't construct a rational decision theory that applies to decision problems that don't satisfy Aesop's principle. Richard Jeffrey (1965) offers a theory in which your own choice of action may provide evidence about the choices made by other people. Ed Karni (1985) offers a theory in which the consequences may be inescapably state-dependent. I am hesitant about such theories because I don't know how to evaluate their basic assumptions.

1.5 Revealed Preference

The theory of revealed preference goes back a long way. Thomas Hobbes (1986), for example, sometimes writes as though choosing a when b is available is unproblematically the same as liking a better than b. However, its modern incarnation is usually attributed to the economist Paul Samuelson (1947), although the basic idea goes back at least as far as Frank Ramsey (1931).

So much confusion surrounds the simple idea of revealed preference that it is worth reviewing its history briefly. The story begins with Jeremy Bentham's adoption of the term *utility* as a measure of the pleasure or pain a person feels as a result of a decision being made. Perhaps he thought that some kind of metering device might eventually be wired into Pandora's brain that would show how many units of utility (utils) she was experiencing. This is a less bold hypothesis than it may once have seemed, since we now know that rats will pull a lever that activates an electrode embedded in a pleasure center in their brains in preference to anything else whatever—including sex and food. It is therefore unsurprising that a modern school of behavioral economists have reverted to this classical understanding of the nature of utility. A more specialized group devote their attention specifically to what they call happiness studies. However, the theory of revealed preference remains the orthodoxy in economic theory (Mas-Collel et al. 1995).

1.5.1 Freeing Economics from Psychology

Economists after Bentham became increasingly uncomfortable, not only with the naive hypothesis that our brains are machines for generating utility, but with all attempts to base economics on psychological foundations. The theory of revealed preference therefore makes a virtue of assuming *nothing whatever* about the psychological causes of our choice behavior.

This doesn't mean that economists believe that our choice behavior isn't caused by what goes on in our heads. Adopting the theory of revealed preference doesn't entail abandoning the principle that reason is the slave of the passions. Studies of brain-damaged people show that when our capacity for emotional response is impaired, we also lose our capacity to make coherent decisions (Damasio 1994). Nor is there any suggestion that we all make decisions in the same way. The theory of revealed preference accepts that some people are clever, and others are stupid; that some care only about money, and others just want to stay out of jail. Nor does the theory insist that people are selfish, as its critics mischievously maintain. It has no difficulty in modeling the kind of saintly folk who would sell the shirt off their back rather than see a baby cry.

Modern decision theory succeeds in accommodating the infinite variety of the human race within a single theory simply by denying itself the luxury of speculating about what is going on inside someone's head. Instead, it pays attention only to what people do. It assumes that we

already know what people choose in some situations, and uses this data to deduce what they will choose in other situations.

For example, Pandora may buy a bundle of goods on each of her weekly visits to the supermarket. Since her household budget and the supermarket prices vary from week to week, the bundle she purchases isn't always the same. However, after observing her shopping behavior for some time, one can make an educated guess about what she will buy next week, once one knows what the prices will be, and how much she will have to spend.

Stability. In making such inferences, two assumptions are implicitly understood. The first is that Pandora's choice behavior is *stable*. We won't be able to predict what she will buy next week if something happens today that makes our data irrelevant. If Pandora loses her heart to a football star, who knows how this might affect her shopping behavior? Perhaps she will buy no pizza at all, and start filling her shopping basket with deodorant instead.

Amartya Sen (1993) offers a rather different example based on the observation that people never take the last apple from a bowl, but will take an apple from the bowl when it contains two apples. Are their preferences over apples therefore unstable, in that sometimes they like apples and sometimes they don't?

Sen's example underlines the care that Pandora must exercise in formulating her decision problem. The people in Sen's story inhabit some last bastion of civilization where Miss Manners still reigns supreme—and this is relevant when Pandora comes to model her problem. Her belief space B must allow her to recognize that she is taking an apple from a bowl in a society that subscribes to the social values of Miss Manners rather than those of Homer Simpson. Her consequence space C must register that the subtle social punishments that her fellows will inflict on her for deviating from their social mores in taking the last apple from a bowl are at least as important to her as whether she gets to eat an apple right now. Pandora's apparent violation of the stability requirement of revealed-preference theory then ceases to bite. Her choice behavior reveals that she likes apples enough to take one when no breach of etiquette is involved, but not otherwise.

1.5.2 Consistency

The second implicit assumption required by the theory of revealed preference is that Pandora's choice behavior must be *consistent*. We certainly won't be able to predict what she will do next if she just picks items off supermarket shelves at random, whether or not they are good value or

satisfy her needs. We are therefore back with the question: What criteria determine whether Pandora's choice behavior is consistent?

Strict preferences. It is usual to write $a \prec b$ to mean that Pandora likes b strictly more than a. Such a strict preference relation is said to be consistent if it is both asymmetric and transitive.

A preference relation $\prec$ is transitive if

$$a \prec b \text{ and } b \prec c \quad \text{implies} \quad a \prec c. \tag{1.1}$$

It is only when transitivity holds that we can describe Pandora's preferences by simply writing $a \prec b \prec c$. Without transitivity, this information wouldn't imply that $a \prec c$.

A preference relation $\prec$ is asymmetric if we don't allow both $a \prec b$ and $b \prec a$. It represents a full set of strict preferences on a set X if we insist that either $a \prec b$ or $b \prec a$ must always hold for any a and b in X that aren't equal.[4]

Given a set of full and consistent strict preferences for Pandora over a finite set X, we can predict the unique alternative $x = \gamma(A)$ that she will choose from each subset A of X. We simply identify $\gamma(A)$ with the alternative x in A for which $a \prec x$ for all other alternatives a in A.

Reversing the logic. Instead of constructing a choice function γ from a given preference relation $\prec$, a theory of revealed preference needs to construct a preference relation from a given choice function.

If we hope to end up with a *strict* preference relation, the choice function γ will need to assign a *unique* optimal outcome $\gamma(A)$ to each subset A of a given finite set X. We can then define $\prec$ by saying that Pandora prefers b to a if and only if she chooses b from the set $\{a, b\}$. That is to say:

$$a \prec b \quad \text{if and only if} \quad b = \gamma\{a, b\}. \tag{1.2}$$

A strict reference relation constructed in this way is automatically asymmetric, but it need not be transitive. To obtain a consistent preference relation, we need the choice function to be consistent in some sense.

Independence of Irrelevant Alternatives. There is a large economic literature (Richter 1988) on consistent choice under market conditions which I propose to ignore. In the absence of information about prices and budgets, the theory of revealed preference has much less bite, but is correspondingly simpler. The only consistency requirement we need right

[4] The technical term is *complete* or *total*.

now is the Independence of Irrelevant Alternatives:[5]

> If Pandora sometimes chooses b when a is feasible, then she never chooses a when b is feasible.

The following argument by contradiction shows that this consistency requirement implies that the revealed-preference relation defined by (1.2) is transitive.

If the transitivity requirement (1.1) is false for some a, b, and c, then $b = \gamma\{a, b\}$, $c = \gamma\{b, c\}$, and $a = \gamma\{c, a\}$. The fact that $b = \gamma\{a, b\}$ shows that b is sometimes chosen when a is feasible. It follows that a can't be chosen from $\{a, b, c\}$ because b is feasible. Similar arguments show that b and c can't be chosen from $\{a, b, c\}$ either. But Pandora must choose one of the alternatives in $\{a, b, c\}$, and so we have a contradiction.

Another example from Amartya Sen (1993) will help to explain why Pandora may have to work hard to find a formulation of her decision problem in which a consistency requirement like the Independence of Irrelevant Alternatives holds.

A respectable lady is inclined to accept an invitation to take tea until she is told that she will also have the opportunity to snort cocaine. She thereby falls foul of the Independence of Irrelevant Alternatives. There are no circumstances in which she would choose to snort cocaine, and so removing the option from her feasible set should make no difference to what she regards as optimal.

The reason that she violates the Independence of Irrelevant Alternatives without our finding her behavior unreasonable is that snorting cocaine isn't an *irrelevant* alternative for her. The fact that cocaine is on the menu changes her beliefs about the kind of person she is likely to meet if she accepts the invitation. If we want to apply the theory of revealed preference to her behavior, we must therefore find a way to formulate her decision problem in which no such hidden relationships link the actions available in her feasible set A with either her beliefs concerning the states in the set B or the consequences in the set C.

Indifference. The preceding discussion was simplified by only looking at cases in which all of Pandora's preferences are strict. Taking account of the possibility of indifference is a bit of a nuisance, but we can deal with the problem very quickly.

If Pandora is sometimes indifferent, it is necessary to abandon the assumption that $\prec$ fully describes her preferences over X. However, we

[5] The Independence of Irrelevant Alternatives is used here in the original sense of Nash (1950). It is unfortunate that Arrow (1963) borrowed the same terminology for a related but different idea.

can define a full relation $\preceq$ on X by

$$a \preceq b \quad \text{if and only if} \quad \text{not}(b \prec a).$$

We say that $\preceq$ is a weak preference relation if it is transitive.[6] We don't need to postulate that the strict preference relation $\prec$ is transitive separately, because this follows from our other assumptions. If $\preceq$ is a weak preference relation, we write $a \sim b$ if and only if $a \preceq b$ and $b \preceq a$. Pandora is then said to be indifferent between a and b.

A weak preference relation $\preceq$ on X doesn't necessarily determine a unique choice for Pandora in each subset A of X because she may be indifferent among several alternatives, all of which are optimal. So the notation $\gamma(A)$ now has to denote the choice *set* consisting of all x in A for which $a \preceq x$ for all a in A. Pandora is assumed to regard each alternative in $\gamma(A)$ as an equally good solution of her decision problem.

When we reverse the logic by seeking to deduce a weak preference relation $\preceq$ from a choice function γ, we must therefore allow the value of $\gamma(A)$ to be any nonempty subset of A. We can still define $\prec$ as in (1.2), but now

$$a \sim b \quad \text{if and only if} \quad \{a, b\} = \gamma\{a, b\}.$$

If the relation $\preceq$ constructed in this way is to be a weak preference relation, we need to impose some consistency requirement on Pandora's choices to ensure that $\preceq$ is transitive. It suffices to strengthen the Independence of Irrelevant Alternatives to Houthakker's axiom:[7]

> If Pandora sometimes includes b in her choice set when a is feasible, then she never includes a in her choice set when b is feasible without including b as well.

1.6 Rationality and Evolution

Evolution is about the survival of the fittest. Entities that promote their fitness consistently will therefore survive at the expense of those that promote their fitness only intermittently. When biological evolution has had a sufficiently long time to operate, it is therefore likely that each relevant locus on a chromosome will be occupied by the gene with maximal fitness. Since a gene is just a molecule, it can't *choose* to maximize its fitness, but evolution makes it seem as though it had. This is a valuable insight, because it allows biologists to use rationality considerations

[6] Which means that (1.1) holds with $\prec$ replaced by $\preceq$.

[7] I follow David Kreps (1988) in attributing the axiom to Houtthaker, but it seems to originate with Arrow (1959).

to predict the outcome of an evolutionary process, without needing to follow each complicated twist and turn that the process might take.

When appealing to rationality in such an evolutionary context, we say that we are seeking an explanation in terms of *ultimate* causes rather than *proximate* causes. Why, for example, do songbirds sing in the early spring? The proximate cause is long and difficult. This molecule knocked against that molecule. This chemical reaction is catalyzed by that enzyme. But the ultimate cause is that the birds are signalling territorial claims to each other in order to avoid unnecessary conflict. They neither know nor care that this behavior is rational. They just do what they do. But the net effect of an immensely complicated evolutionary process is that songbirds behave *as though* they had rationally chosen to maximize their fitness.

Laboratory experiments on pigeons show that they sometimes honor various consistency requirements of rational choice theory better than humans (Kagel, Battalio, and Green 1995). We don't know the proximate explanation. Who knows what goes on inside the mind of a pigeon? Who knows what goes on in the minds of stockbrokers for that matter? But we don't need to assume that pigeons or stockbrokers are thinking rationally because we see them behaving rationally. We can appeal to the ultimate explanations offered by evolutionary theory.

People naturally think of biology when such appeals to evolution are made, but I follow the practice in economics of using the term more widely. After Alchian (see Lott 1997), it is common to argue that the forces of social or economic evolution will tend to eliminate stockbrokers who don't consistently seek to maximize profit. It is therefore unsurprising that evolutionary arguments are sometimes marshaled in defense of rationality concepts. The money pump argument is a good example.

Money pumps. Why should we expect Pandora to reveal transitive preferences? The money pump argument says that if she doesn't, other people will be able to make money out of her.

Suppose that Pandora reveals the intransitive preferences

$$\text{apple} \prec \text{orange} \prec \text{fig} \prec \text{apple}$$

when making pairwise comparisons. A trader now gives her an apple. He then offers to exchange the apple for an orange, provided she pays him a penny. Since her preference for the orange is strict, she agrees. The trader now offers to exchange the orange for a fig, provided she pays him a penny. When she agrees, the trader offers to exchange the fig for an apple, provided she pays him a penny.

If Pandora's preferences are stable, this trading cycle can be repeated until Pandora is broke. The inference is that nobody with intransitive preferences will be able to survive in a market context.

1.7 Utility

In the theory of revealed preference, utility functions are no more than a mathematical device introduced to help solve choice problems. A preference relation $\preceq$ is represented by a real-valued utility function u if and only if

$$u(a) \leqslant u(b) \quad \text{if and only if} \quad a \preceq b.$$

Finding an optimal x then reduces to solving the maximization problem:

$$u(x) = \max_{a \in A} u(a), \tag{1.3}$$

for which many mathematical techniques are available.

Constructing utility functions. Suppose that Pandora's choice behavior reveals that she has consistent preferences over the five alternatives a, b, c, d, and e. Her revealed preferences are

$$a \prec b \sim c \prec d \prec e.$$

It is easy to find a utility function U that represents Pandora's preferences. She regards the alternatives a and e as the worst and the best available. We therefore set $U(a) = 0$ and $U(e) = 1$. We next pick any alternative intermediate between the worst and the best alternative, and take its utility to be $\frac{1}{2}$. In Pandora's case, b is an alternative intermediate between a and e, and so we set $U(b) = \frac{1}{2}$. Since $b \sim c$, we must also set $U(c) = \frac{1}{2}$. Only the alternative d remains. This is intermediate between c and e, and so we set $U(d) = \frac{3}{4}$, because $\frac{3}{4}$ is intermediate between $U(c) = \frac{1}{2}$ and $U(e) = 1$.

The utilities we have assigned to alternatives are ranked in the same way as the alternatives themselves. In making choices, Pandora therefore behaves *as though* she were maximizing the value of U. But she also behaves as though she were maximizing the values of the alternative utility functions V and W. There are also many other ways that we could have assigned utilities to the alternatives in a manner consistent with Pandora's preferences. In the theory of revealed preference, the only criterion that is relevant when picking one of the infinity of utility functions that represent a given preference relation is that of mathematical convenience.

a	a	b	c	d	e
$U(a)$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$V(a)$	-1	1	1	2	3
$W(a)$	-8	0	0	1	8

Figure 1.2. Constructing utility functions. The method always works for a consistent preference relation defined over a finite set of alternatives, because there is always another real number between any pair of real numbers.

1.7.1 Cardinal versus Ordinal

The distinction between ordinal and cardinal notions of utility is left over from a controversy of the early twentieth century. Jeremy Bentham's identification of utility with pleasure or pain was the orthodox position among early Victorian economists, but a generation of reforming economists led by Stanley Jevons (1871) showed that one can often dispense with appeals to utility altogether by carefully considering what happens "at the margin." In a manner familiar to historians of thought, the generation following the marginalist revolution then poured scorn on what they saw as the errors of their Victorian forebears. The jeremiads of the economist Lionel Robbins (1938) are still quoted by modern critics who are unaware that modern utility theory has totally abandoned the shaky psychological foundations proposed by Bentham and his followers.

Robbins saw no harm in the use of ordinal utility functions as a mathematical convenience. These are utility functions constructed like those of figure 1.2 so that only the *ranking* of their values has any significance. For example, the fact that $U(e) - U(d) = U(d) - U(c)$ doesn't tell us that Pandora would be as willing to swap d for e as she would be to swap c for d. After all, we were free to assign any value to $U(d)$ between $\frac{1}{2}$ and 1. The fact that only the ranking of alternatives is preserved in an ordinal utility function is expressed mathematically by saying that any ordinal utility function is a strictly increasing transformation of any other ordinal utility function that represents the same preference relation. For example, $V(a) = -1 + 4U(a)$ and $W(a) = \{V(a) - 1\}^3$.

One might think of the utility scale created by an ordinal utility function as being marked on an elastic measuring tape. No matter how one squeezes or stretches the tape, the result will still be an ordinal utility

scale. A cardinal utility function is one that creates a scale like temperature. One is free to choose the zero and the unit on such a scale, but then there is no more room for maneuver. In mathematical terms, any cardinal utility function is only a strictly increasing *affine* transformation of any other cardinal utility function that represents the same preference relation.[8]

Robbins drew the line at cardinal utility functions. It is ironic that he was still denouncing them as intrinsically meaningless long after the moment at which John Von Neumann explained to Oskar Morgenstern how to make sense of them in risky situations. However, this is a story that must wait until section 3.4.

The immediate point is that Robbins was perfectly right to insist that the mere fact that we might somehow be gifted with cardinal utility functions doesn't imply that we can necessarily compare Pandora's utils meaningfully with those of other people. To do so would be like thinking that a room in which the temperature is 32 °Fahrenheit must be warmer than a room in which the temperature is 31 °Celsius. If we want to use utility functions to make interpersonal comparisons of welfare, we certainly need a theory that generates cardinal utility functions rather than the ordinal utility functions considered in this chapter, but we also need a lot more input concerning the social context (section 4.4).

1.7.2 Infinite Sets

It is sometimes argued that infinite sets don't actually exist, and so we need not make our lives complicated by worrying about them. But this argument completely misses the point. Infinite sets aren't introduced into models for metaphysical reasons. They are worth worrying about because infinite models are often much *simpler* than the finite models for which they serve as idealizations.

Finding a utility representation of a consistent preference relation defined on an infinite set X is no problem when X is countable.[9] The method of section 1.6 works equally well in this case, provided we allow the utility function to take values outside the range determined by the first pair of alternatives that we choose to consider. If we like, the values of the utility function so constructed can be made to lie between any two predetermined bounds.

[8] The equation $y = Ax + B$ defines an affine transformation from the real numbers to the real numbers. It is strictly increasing if and only if $A > 0$. Thus V can be obtained by applying the strictly increasing affine transformation $y = 4x - 1$ to U.

[9] This means that the members of X can be arranged in a sequence and so be counted. The set of all n-tuples of rational numbers is a countable dense subset of the uncountable set $\mathbb{R}^n$ of all n-tuples of real numbers.

Given a consistent preference relation on a set X with a countable dense subset Y, we can first construct a utility representation on Y as above. The representation can then be extended to the whole of X by continuity, provided that all the sets $\{a \in X : a \preceq b\}$ and $\{a \in X : a \succeq b\}$ are closed for all b in X. The resulting utility function will then be continuous, which guarantees that the maximization problem (1.3) has a solution when the feasible set A is compact.[10]

Lexicographic preferences. It isn't true that all consistent preference relations on infinite sets have a utility representation. For example, in the torture example of the next section, Pandora might care primarily about the intensity of pain she must endure, taking account of the time it must be endured only when comparing two situations in which the intensity of pain is the same. It is impossible to represent such a lexicographic preference with a utility function.[11]

Any such representation would assign an open interval of real numbers to each vertical line in figure 1.3. But there are an uncountable number of such verticals, and any collection of nonoverlapping open intervals must be countable (because we can label each such open interval with one of the rational numbers it contains).

1.8 Challenging Transitivity

This section reviews two of the many arguments that continue to be directed against transitivity by philosophers.

Paradox of preference. This example is taken from a recent compendium of paradoxes put together by Clark (2002). You would choose to fly a glider for the first time accompanied by an experienced instructor rather than try out a racing car. But you would choose the racing car rather than flying the glider solo. However, your machismo requires that you choose flying the glider solo to flying with an instructor if that is the only choice on offer.

As with Sen's examples, the problem hasn't been adequately formulated. At the very least, it is necessary to distinguish between doing something without loss of machismo and doing it with loss of machismo.

[10] To say that Y is dense in X means that we can approximate any point in X as closely as we like by points in Y. To say that a set is closed means that it contains all its boundary points. To say that a set in $\mathbb{R}^n$ is compact means that it is closed and bounded.

[11] Lexicographic means alphabetical. If Pandora had lexicographic preferences over words, then she would prefer whichever of two words came first in the dictionary.

Achilles and the tortoise. In arguing that the relationship "all things considered better than" need not be transitive, Stuart Rachels (1998, 2001) and Larry Temkin (1996) make three claims about human experience:[12]

Claim 1. Any unpleasant experience, no matter what its intensity and duration, is better than a slightly less intense experience that lasts much longer.

Claim 2. There is a finely distinguishable range of unpleasant experiences ranging in intensity from mild discomfort to extreme agony.

Claim 3. No matter how long it must be endured, mild discomfort is preferable to extreme agony for a significant amount of time.

These three claims are invoked to justify a story, which begins with our being invited to consider two extreme alternatives. In the first, Pandora suffers an excruciating torture for a short period of time. In the second, she endures a mild pain for a very long time. Between these two extremes lie a chain of alternatives in which the pain keeps being reduced very slightly but the time it must be endured is prolonged so that Pandora never regards the next alternative as better than its predecessor. But Pandora prefers the final alternative to the initial alternative, and so her preferences go round in a circle, which transitivity doesn't allow.

The argument can be shown to be wrong simply by considering the special case in which Pandora's preferences are determined by the utility function

$$u(p,t) = -pt/(1+t),$$

in which p is the intensity of pain and t the length of time it must be endured. The preferences that Pandora reveals are necessarily transitive, because this is always true of preferences obtained from a utility function. However, Pandora nevertheless satisfies all three of the claims said to imply that her preferences must sometimes be intransitive.

Figure 1.3 shows how utility functions are often illustrated in diagrams by drawing indifference curves along which Pandora's utility is constant. The little arrows show the direction of Pandora's preference at any pair (p,t) representing an intensity p of pain endured for a period of length t. The chain of alternatives in the argument are shown as the sequence $(p_1,t_1),(p_2,t_2),(p_3,t_3),\ldots$. The figure makes it clear that we aren't entitled to assume that the intensity of pain in this sequence can be brought as close to zero as we choose. To assume otherwise is to fall prey to the paradox of Achilles and the tortoise.

[12] See Binmore and Voorhoeve (2003, 2006) and Voorhoeve (2007) for a less terse account. Quinn's (1990) paradox, which is similar but substitutes the Sorites paradox for Zeno's paradox, is also considered.

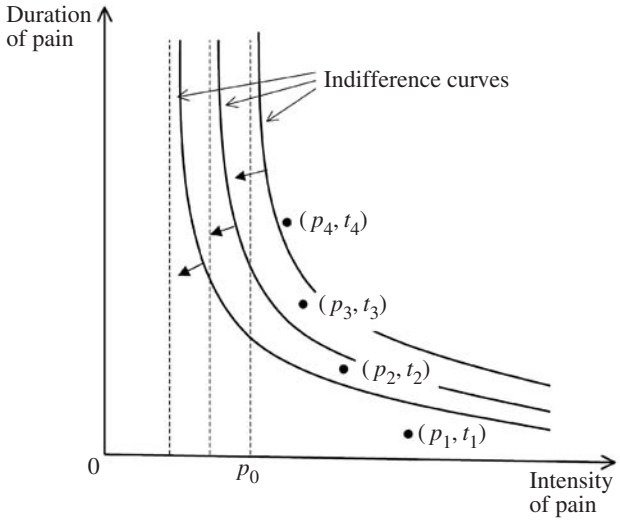

Figure 1.3. Achilles and the tortoise. Pandora likes it less and less as she progresses through the sequence (p_n, t_n), but she never gets to a stage where the pain is negligible, because $p_n > p_0$ for all values of n.

My own view is that such arguments fail to take proper account of the proviso that "all things are to be considered." Pandora should also consider what she would choose if offered feasible sets containing *three* alternatives. If she doesn't violate Houthakker's axiom, the preferences she reveals will then necessarily be transitive.

1.9 Causal Utility Fallacy

The fact that utility used to mean one thing and now means another is understandably the cause of widespread confusion. Bentham himself suggested substituting the word *felicity* for his notion of utility as pleasure or pain, but I suppose it is hopeless to propose that his suggestion be taken up at this late date. The chief misunderstanding that arises from confusing a utility derived from revealed-preference theory with a Benthamite felicity is that the former offers no explanation of why Pandora should choose one thing rather than another, while the latter does.

In revealed-preference theory, it isn't true that Pandora chooses b rather than a *because* the utility of b exceeds the utility of a. This is the Causal Utility Fallacy. It isn't even true that Pandora chooses b rather than a because she prefers b to a. On the contrary, it is because Pandora chooses b rather than a that we say that Pandora prefers b to a, and assign b a larger utility.

The price of abandoning psychology for revealed-preference theory is therefore high. We have to give up any pretension to be offering a causal explanation of Pandora's choice behavior in favor of an account that is merely a description of the choice behavior of someone who chooses consistently. Our reward is that we end up with a theory that is hard to criticize because it has little substantive content.

Notice that Pandora might be choosing consistently because she is seeking to maximize the money in her pocket, or consciously pursuing some other end. In an evolutionary context, Pandora might be an animal who isn't choosing consciously at all, but who makes consistent choices because animals who failed to promote their fitness consistently have long since been eliminated. Neither of these possibilities is denied by the theory of revealed preference, which is entirely neutral about *why* Pandora makes rational decisions.

In such cases, we can simply identify utility with money or fitness without any hassle. It is only when no obvious maximand is in sight that the theory of revealed preference comes into its own. It tells us that if Pandora's choice behavior is consistent, then we can model her as maximizing some abstract quantity called utility. She may not be aware that she is behaving as though maximizing the utility function we have invented for her, but she behaves as though she is a maximizer of utility nevertheless.

Neoclassical economics. The theory of revealed preference is a child of the marginalist revolution. As such, it is an official doctrine of neoclassical economics, enshrined in all respectable textbooks.

However, the kind of critic who thinks that economists are mean-minded, money-grubbing misfits commonly ignores the official doctrine in favor of a straw man that is easier to knock down. It is said that neoclassical economics is based on the principle that people are selfish. Henrich et al. (2004) even invent a "selfishness axiom" that supposedly lies at the heart of economics. Sometimes it is said that economists necessarily believe that people care only about money, so that utility theory reduces to a crude identification of utils with dollars.

It is true that economists are more cynical about human behavior than the general run of humanity. I became markedly more cynical myself after becoming a business consultant to raise money for my research center. It is an unwelcome fact that many people do behave like mean-minded, money-grubbing misfits in the business world. Nor is it true that the laboratory experiments of behavioral economists show that ordinary people are always better behaved. After ten trials or so, nearly all subjects

end up defecting in the Prisoners' Dilemma (Ledyard 1995). However, the empirical facts on the ground are irrelevant to the immediate issue.

It isn't true that it is axiomatic in economic theory that people are selfish. I suspect that this widespread misunderstanding arises because people think that rational agents must act out of self-interest because they maximize their own utility functions rather than some social objective function. But to make this claim is to fail to understand the theory of revealed preference. Whatever the explanation of Pandora's behavior, if she acts consistently, she will act as though maximizing a utility function tailored to her own behavior. Tradition is doubtless right that St Francis of Assisi consistently promoted the interests of the poor and sick. He therefore behaved as though maximizing a utility function that took account of their welfare. But nobody would want to say that St Francis was selfish because the particular form of his utility function was idiosyncratic to him.

In the past, I have followed certain philosophers in saying that people act in their own *enlightened* self-interest when they act consistently, but I now see that this was a mistake. Any kind of language that admits the Causal Utility Fallacy as a possibility is best avoided.

Preference satisfaction. Most accounts of rational decision theory in the philosophical literature fall headlong into the Causal Utility Fallacy.[13] For example, Gauthier (1993, p. 186) tells us that "only utilities provide reasons for acting." Accounts that avoid this mistake usually follow A. J. Ayer in talking about "preference satisfaction." Utility functions are indeed constructed from preferences, but preferences themselves aren't primitives in the theory of revealed preference. The primitives of the theory are the *choices* that Pandora is seen (or hypothesized) to make.

Rational choice theory. There is an ongoing debate in political science about the extent to which what they call rational choice theory can sensibly be applied. Neither side in this often heated debate will find much comfort in the views expressed in this chapter. The issue isn't whether people are capable of the superhuman feats of calculation that would be necessary to consciously maximize a utility function in most political contexts. After all, spiders and fish can't be said to be conscious at all, but evolutionary biologists nevertheless sometimes succeed in modeling them as maximizers of biological fitness. As in philosophy, both the

[13] Simon Blackburn's (1998, chapter 6) *Ruling Passions* is an interesting philosophical text that gives an accurate account of revealed-preference theory. Like Gibbard (1990), Blackburn seeks proximate causes of our moral behavior rooted in our emotional responses. My own complementary approach looks for ultimate causes in our evolutionary history (Binmore 2005).

proponents and the opponents of rational choice theory need to learn that their theory is based on consistency of *choice* (for whatever reason) rather than on conscious preference satisfaction.

1.10 Positive and Normative

What is rationality good for? The standard answer is that it is good for working out the ideal means of achieving whatever your ends may be. Economists traditionally offer another answer. They attribute rationality to the agents in their models when seeking to predict how ordinary people will actually behave in real situations. Both answers deserve some elaboration.

Normative applications. A theory is normative or prescriptive if it says what ought to happen. It is positive or descriptive if it predicts what will actually happen.

Jeremy Bentham's approach to utility was unashamedly normative. He tells us that, in borrowing the idea of utility from David Hume, his own contribution was to switch the interpretation from positive to normative. Many pages of his writings are devoted to lists of all the many factors that might make Pandora happy or sad. But the modern theory of rational decisions doesn't presume to tell you what your aims should be. Nor is it very specific about the means to be adopted to achieve whatever your ends may be. Indeed, since the modern theory can be regarded as a positive theory of behavior for an idealized world of rational agents, it needs to be explained why it has a normative role at all.

Pandora uses the theory of revealed preference normatively when she revises her attitudes to the world after discovering that her current attitudes would lead her to make choices in some situations that are inconsistent with the choices she would make in other situations.

A famous example arose when Leonard Savage was entertained to dinner by the French economist Maurice Allais. Allais asked Savage how he would choose in some difficult-to-assess situations (section 3.5). When Savage gave inconsistent answers, Allais triumphantly declared that even Savage didn't believe his own theory. Savage's response was to say that he had made a mistake. Now that he understood that his initial snap responses to Allais' questions had proved to generate inconsistencies, he would revise his planned choices until they became consistent.

One doesn't need to dine with Nobel laureates in Paris to encounter situations in which people use their rationality in revising snap judgments. I sometimes ask finance experts whether they prefer 96×69 dollars to 87×78 dollars. If given no time to think, most say the former. But when

it is pointed out that $96 \times 69 = 6,624$ and $87 \times 78 = 6,786$, they always change their minds. An anecdote from Amos Tversky (2003) makes a similar point. In a laboratory experiment, many of his subjects made intransitive choices. When this was pointed out, a common response was to claim that his records were mistaken—the implication being that they wouldn't have made intransitive choices if they had realized they were doing so.

Positive applications. Rational decision theory isn't very good at predicting the choice behavior of inexperienced or unmotivated people. For example, laboratory experiments show that most people choose option (A) in problem 1 below and option (B) in problem 2, although the two problems differ only in how they are framed. Their choices are therefore unstable to irrelevant framing effects.

Problem 1. You are given $200 and then must choose between:

(A) $50 extra for sure;

(B) $200 extra with probability 25%.

Problem 2. You are given $400 and then must choose between:

(A) $150 less for sure;

(B) $200 less with probability 75%.

It is for this kind of reason that professors of marketing teach their students to laugh at economic consumer theory, which assumes that shoppers behave rationally when buying spaghetti or toothpaste.

However, there is a good deal of evidence that adequately motivated people sometimes can and do learn to choose rationally if they are put in similar situations repeatedly and provided with meaningful feedback on the outcome of their previous decisions (Binmore 2007a). Sometimes the learning is conscious, as when insurance companies systematically seek to maximize their long-term average profit. But most learning is unconscious. Like the squirrels who always find a way to get round the increasingly elaborate devices with which I have tried to defend my bird feeder, people mostly learn by trial-and-error. However, it is necessary to face up to the fact that the laboratory evidence suggests that humans find trial-and-error learning especially difficult when the feedback from our choices is confused by chance events, as will be the case in most of this book.

Fortunately, we don't just learn by trial-and-error. We also learn from books. Just as it is easier to predict how educated kids will do arithmetic, so the spread of decision theory into our universities and business

schools will eventually make it easier to predict how decisions get made in economic life. If Pandora knows that $96 \times 69 = 6,624$ and $87 \times 78 = 6,786$, she won't make the mistake of choosing 96×69 dollars over 87×78 dollars. Once Allais had taught Savage that his choice behavior was inconsistent, Savage changed his mind about how to choose.

In brief, rational decision theory is only a useful positive tool when the conditions are favorable. Economists sometimes manage to convince themselves that the theory always applies to everything, but such enthusiasm succeeds only in providing ammunition for skeptics looking for an excuse to junk the theory altogether.