1

Decision Theory and Human Behavior

People are not logical. They are *psycho*logical.
Anonymous

People often make mistakes in their maths. This does not mean that we should abandon arithmetic.
Jack Hirshleifer

Decision theory is the analysis of the behavior of an individual facing nonstrategic uncertainty—that is, uncertainty that is due to what we term "Nature" (a stochastic natural event such as a coin flip, seasonal crop loss, personal illness, and the like) or, if other individuals are involved, their behavior is treated as a statistical distribution known to the decision maker. Decision theory depends on probability theory, which was developed in the seventeenth and eighteenth centuries by such notables as Blaise Pascal, Daniel Bernoulli, and Thomas Bayes.

A *rational actor* is an individual with *consistent preferences* (§1.1). A rational actor need not be selfish. Indeed, if rationality implied selfishness, the only rational individuals would be sociopaths. Beliefs, called *subjective priors* in decision theory, logically stand between choices and payoffs. Beliefs are primitive data for the rational actor model. In fact, beliefs are the product of social processes and are shared among individuals. To stress the importance of beliefs in modeling choice, I often describe the rational actor model as the *beliefs, preferences and constraints* model, or the *BPC model*. The BPC terminology has the added attraction of avoiding the confusing and value-laden term "rational."

The BPC model requires only preference consistency, which can be defended on basic evolutionary grounds. While there are eminent critics of preference consistency, their claims are valid in only a few narrow areas. Because preference consistency does not presuppose unlimited information-processing capacities and perfect knowledge, even *bounded rationality* (Simon 1982) is consistent with the BPC model.[1] Because one cannot do behavioral game theory, by which I mean the application of game theory to the experimental study of human behavior, without assuming preference consistency, we must accept this axiom to avoid the analytical weaknesses of the behavioral disciplines that reject the BPC model, including psychology, anthropology, and sociology (see chapter 12).

Behavioral decision theorists have argued that there are important areas in which individuals appear to have inconsistent preferences. Except when individuals do not know their own preferences, this is a conceptual error based on a misspecification of the decision maker's preference function. We show in this chapter that, assuming individuals know their preferences, adding information concerning the current state of the individual to the choice space eliminates preference inconsistency. Moreover, this addition is completely reasonable because preference functions do not make any sense unless we include information about the decision maker's current state. When we are hungry, scared, sleepy, or sexually deprived, our preference ordering adjusts accordingly. The idea that we should have a utility function that does not depend on our current wealth, the current time, or our current strategic circumstances is also not plausible. Traditional decision theory ignores the individual's current state, but this is just an oversight that behavioral decision theory has brought to our attention.

Compelling experiments in behavioral decision theory show that humans violate the principle of expected utility in systematic ways (§1.7). Again, is must be stressed that this does *not* imply that humans violate preference consistency over the appropriate choice space but rather that they have incorrect beliefs deriving from what might be termed "folk probability theory" and make systematic *performance errors* in important cases (Levy 2008).

To understand why this is so, we begin by noting that, with the exception of hyperbolic discounting when time is involved (§1.4), there are no reported failures of the expected utility theorem in nonhumans, and there are some extremely beautiful examples of its satisfaction (Real 1991). Moreover, territoriality in many species is an indication of loss aversion (Chapter 11). The difference between humans and other animals is that the latter are tested in *real life*, or in elaborate simulations of real life, as in Leslie Real's work with bumblebees (1991), where subject bumblebees are re-

[1] Indeed, it can be shown (Zambrano 2005) that every boundedly rational individual is a fully rational individual subject to an appropriate set of Bayesian priors concerning the state of nature.

leased into elaborate spatial models of flowerbeds. Humans, by contrast, are tested using imperfect *analytical models* of real-life lotteries. While it is important to know how humans choose in such situations, there is certainly no guarantee they will make the same choices in the real-life situation and in the situation analytically generated to represent it. Evolutionary game theory is based on the observation that individuals are more likely to adopt behaviors that appear to be successful for others. A heuristic that says "adopt risk profiles that appear to have been successful to others" may lead to preference consistency even when individuals are incapable of evaluating analytically presented lotteries in the laboratory.

In addition to the explanatory success of theories based on the BPC model, supporting evidence from contemporary neuroscience suggests that expected utility maximization is not simply an "as if" story. In fact, the brain's neural circuitry actually makes choices by internally representing the payoffs of various alternatives as neural firing rates and choosing a maximal such rate (Shizgal 1999; Glimcher 2003; Glimcher and Rustichini 2004; Glimcher, Dorris, and Bayer 2005). Neuroscientists increasingly find that an aggregate decision making process in the brain synthesizes all available information into a single unitary value (Parker and Newsome 1998; Schall and Thompson 1999). Indeed, when animals are tested in a repeated trial setting with variable rewards, dopamine neurons appear to encode the difference between the reward that the animal expected to receive and the reward that the animal actually received on a particular trial (Schultz, Dayan, and Montague 1997; Sutton and Barto 2000), an evaluation mechanism that enhances the environmental sensitivity of the animal's decision making system. This error prediction mechanism has the drawback of seeking only local optima (Sugrue, Corrado, and Newsome 2005). Montague and Berns (2002) address this problem, showing that the orbitofrontal cortex and striatum contain a mechanism for more global predictions that include risk assessment and discounting of future rewards. Their data suggest a decision-making model that is analogous to the famous Black-Scholes options-pricing equation (Black and Scholes 1973).

The existence of an integrated decision-making apparatus in the human brain itself is predicted by evolutionary theory. The fitness of an organism depends on how effectively it make choices in an uncertain and varying environment. Effective choice must be a function of the organism's state of knowledge, which consists of the information supplied by the sensory inputs that monitor the organism's internal states and its external environment. In relatively simple organisms, the choice environment is primitive and is distributed in a decentralized manner over sensory inputs. But in three separate groups of animals, craniates (vertebrates and related creatures), arthropods (including insects, spiders, and crustaceans), and cephalopods (squid, octopuses, and other mollusks), a central nervous system with a brain (a centrally located decision-making and control apparatus) evolved. The phylogenetic tree of vertebrates exhibits increasing complexity through time and increasing metabolic and morphological costs of maintaining brain activity. Thus, *the brain evolved because larger and more complex brains, despite their costs, enhanced the fitness of their carriers*. Brains therefore are ineluctably structured to make consistent choices in the face of the various constellations of sensory inputs their bearers commonly experience.

Before the contributions of Bernoulli, Savage, von Neumann, and other experts, no creature on Earth knew how to value a lottery. The fact that people do not know how to evaluate abstract lotteries does not mean that they lack consistent preferences over the lotteries that they face in their daily lives.

Despite these provisos, experimental evidence on choice under uncertainty is still of great importance because in the modern world we are increasingly called upon to make such “unnatural” choices based on scientific evidence concerning payoffs and their probabilities.

1.1 Beliefs, Preferences, and Constraints

In this section we develop a set of behavioral properties, among which consistency is the most prominent, that together ensure that we can model agents as maximizers of preferences.

A *binary relation* $\odot_{\mathrm{A}}$ on a set A is a subset of $A \times A$. We usually write the proposition $(x, y)\psi\in \odot_{\mathrm{A}}$ as $x\psi\odot_{\mathrm{A}}\, y$. For instance, the arithmetical operator “less than” ($<$) is a binary relation, where $(x, y) \in <$ is normally written $x < y$.[2] A *preference ordering* $\succeq_{\mathrm{A}}$ on A is a binary relation with the following three properties, which must hold for all $x, y, z\psi\in A$ and any set B:

1. **Complete**: $x\psi\succeq_{\mathrm{A}} y$ or $y\psi\succeq_{\mathrm{A}} x$;
2. **Transitive**: $x\psi\succeq_{\mathrm{A}} y\psi$and $y\psi\succeq_{\mathrm{A}} z$ imply $x \succeq_{\mathrm{A}} z$;

[2] See chapter 14 for the basic mathematical notation used in this book. Additional binary relations over the set **R** of real numbers include $>$, $<$, $\leq$, $=$, $\geq$, and $\neq$, but $+$ is not a binary relation because $x + y$ is not a proposition.

3. **Independent of irrelevant alternatives**: For $x, y \in B$, $x \succeq_B y$ if and only if $x \succeq_A y$.

Because of the third property, we need not specify the choice set and can simply write $x \succeq y$. We also make the behavioral assumption that given any choice set A, the individual chooses an element $x \in A$ such that for all $y \in A$, $x \succeq y$. When $x \succeq y$, we say "x is weakly preferred to y."

The first condition is *completeness*, which implies that any member of A is weakly preferred to itself (for any x in A, $x \succeq x$). In general, we say a binary relation $\odot$ is *reflexive* if, for all x, $x \odot x$. Thus, completeness implies reflexivity. We refer to $\succeq$ as "weak preference" in contrast with "strong preference" $\succ$. We define $x \succ y$ to mean "it is false that $y \succeq x$." We say x and y are *equivalent* if $x \succeq y$ and $y \succeq x$, and we write $x \simeq y$. As an exercise, you may use elementary logic to prove that if $\succeq$ satisfies the completeness condition, then $\succ$ satisfies the following *exclusion* condition: if $x \succ y$, then it is false that $y \succ x$.

The second condition is *transitivity*, which says that $x \succeq y$ and $y \succeq z$ imply $x \succeq z$. It is hard to see how this condition could fail for anything we might like to call a preference ordering.[3] As a exercise, you may show that $x \succ y$ and $y \succeq z$ imply $x \succ z$, and $x \succeq y$ and $y \succ z$ imply $x \succ z$. Similarly, you may use elementary logic to prove that if $\succeq$ satisfies the completeness condition, then $\simeq$ is transitive (i.e., satisfies the transitivity condition).

The third condition, *independence of irrelevant alternatives* (IIA) means that the relative attractiveness of two choices does not depend upon the other choices available to the individual. For instance, suppose an individual generally prefers meat to fish when eating out, but if the restaurant serves lobster, the individual believes the restaurant serves superior fish, and hence prefers fish to meat, even though he never chooses lobster; thus, IIA fails. When IIA fails, it can be restored by suitably refining the choice set. For instance, we can specify two qualities of fish instead of one, in the preceding example. More generally, if the desirability of an outcome x depends on the set A from which it is chosen, we can form a new choice space Ω^*, elements of which are ordered pairs (A, x), where $x \in A \subseteq \Omega$, and restrict choice sets in Ω^* to be subsets of Ω^* all of whose first elements are equal. In this new choice space, IIA is trivially satisfied.

[3] The only plausible model of intransitivity with some empirical support is *regret theory* (Loomes 1988; Sugden 1993). Their analysis applies, however, only to a narrow range of choice situations.

When the preference relation $\succeq$ is complete, transitive, and independent of irrelevant alternatives, we term it *consistent*. If $\succeq$ is a consistent preference relation, then there will always exist a preference function such that the individual behaves as if maximizing this preference function over the set A from which he or she is constrained to choose. Formally, we say that a preference function $u: A \rightarrow \mathbf{R}$ *represents* a binary relation $\succeq$ if, for all $x, y \in A$, $u(x) \geq u(y)$ if and only if $x \succeq y$. We have the following theorem.

THEOREM 1.1 *A binary relation $\succeq$ on the finite set A of payoffs can be represented by a preference function $u{:}A \rightarrow \mathbf{R}$ if and only if $\succeq$ is consistent.*

It is clear that $u(\cdot)$ is not unique, and indeed, we have the following theorem.

THEOREM 1.2 *If $u(\cdot)$ represents the preference relation $\succeq$ and $f(\cdot)$ is a strictly increasing function, then $v(\cdot) = f(u(\cdot))$ also represents $\succeq$. Conversely, if both $u(\cdot)$ and $v(\cdot)$ represent $\succeq$, then there is an increasing function $f(\cdot)$ such that $v(\cdot) = f(u(\cdot))$.*

The first half of the theorem is true because if f is strictly increasing, then $u(x) > u(y)$ implies $v(x) = f(u(x)) > f(u(y)) = v(y)$, and conversely. For the second half, suppose $u(\cdot)$ and $v(\cdot)$ both represent $\succeq$, and for any $y \in \mathbf{R}$ such that $v(x) = y$ for some $x \in X$, let $f(y) = u(v^{-1}(y))$, which is possible because v is an increasing function. Then $f(\cdot)$ is increasing (because it is the composition of two increasing functions) and $f(v(x)) = u(v^{-1}(v(x))) = u(x)$, which proves the theorem. ■

1.2 The Meaning of Rational Action

The origins of the BPC model lie in the eighteenth century research of Jeremy Bentham and Cesare Beccaria. In his *Foundations of Economic Analysis* (1947), economist Paul Samuelson removed the hedonistic assumptions of utility maximization by arguing, as we have in the previous section, that utility maximization presupposes nothing more than transitivity and some harmless technical conditions akin to those specified above.

Rational does not imply self-interested. There is nothing irrational about caring for others, believing in fairness, or sacrificing for a social ideal. Nor do such preferences contradict decision theory. For instance, suppose a man with \$100 is considering how much to consume himself and how much to give to charity. Suppose he faces a tax or subsidy such that for each \$1 he contributes to charity, he is obliged to pay pψdollars. Thus, $p > 1$ψ represents a tax, while $0 < p < 1$ represents a subsidy. We can then treat pψas the *price* of a unit contribution to charity and model the individual as maximizing his utility for personal consumption xψand contributions to charity y, say $u(x, y)$ subject to the budget constraint $x+py = 100$. Clearly, it is perfectly rational for him to choose $y > 0$. Indeed, Andreoni and Miller (2002) have shown that in making choices of this type, consumers behave in the same way as they do when choosing among personal consumption goods; i.e., they satisfy the generalized axiom of revealed preference.

Decision theory does not presuppose that the choices people make are welfare-improving. In fact, people are often slaves to such passions as smoking cigarettes, eating junk food, and engaging in unsafe sex. These behaviors in no way violate preference consistency.

If humans fail to behave as prescribed by decision theory, we need not conclude that they are irrational. In fact, they may simply be ignorant or misinformed. However, if human subjects consistently make intransitive choices over lotteries (e.g., §1.7), then either they do not satisfy the axioms of expected utility theory or they do not know how to evaluate lotteries. The latter is often called *performance error*. Performance error can be reduced or eliminated by formal instruction, so that the experts that society relies upon to make efficient decisions may behave quite rationally even in cases where the average individual violates preference consistency.

1.3 Why Are Preferences Consistent?

Preference consistency flows from evolutionary biology (Robson 1995). Decision theory often applies extremely well to nonhuman species, including insects and plants (Real 1991; Alcock 1993; Kagel, Battalio, and Green 1995). Biologists define the *fitness* of an organism as its expected number of offspring. Assume, for simplicity, asexual reproduction. A maximally fit individual will then produce the maximal expected number of offspring, each of which will inherit the genes for maximal fitness. Thus, fitness maximization is a precondition for evolutionary survival. If organisms maximized fitness directly, the conditions of decision theory would be directly satisfied because we could simply represent the organism's utility function as its fitness.

However, organisms do *not* directly maximize fitness. For instance, moths fly into flames and humans voluntarily limit family size. Rather, organisms have preference orderings that are themselves subject to selection according to their ability to promote fitness (Darwin 1872). We can expect preferences to satisfy the completeness condition because an organism must be able to make a consistent choice in any situation it habitually faces or it will be outcompeted by another whose preference ordering can make such a choice.

Of course, unless the current environment of choice is the same as the historical environment under which the individual's preference system evolved, we would not expect an individual's choices to be fitness-maximizing, or even necessarily welfare-improving.

This biological explanation also suggests how preference consistency might fail in an imperfectly integrated organism. Suppose the organism has three decision centers in its brain, and for any pair of choices, majority rule determines which the organism prefers. Suppose the available choices are A, B, and $C\psi$and the three decision centers have preferences $A \succ B\psi\succ C$, $B\psi\succ C\psi\succ A$, and $C\psi\succ A \succ B$, respectively. Then when offered A or B, the individual chooses A, when offered $B\psi$or C, the individual chooses B, and when offered A and C, the individual chooses C. Thus $A \succ B\psi\succ C\psi\succ A$, and we have intransitivity. Of course, if an objective fitness is associated with each of these choices, Darwinian selection will favor a mutant who suppresses two of the three decision centers or, better yet, integrates them.

1.4 Time Inconsistency

Several human behavior patterns appear to exhibit *weakness of will*, in the sense that if there is a long time period between choosing and experiencing the costs and benefits of the choice, individuals can choose wisely, but when costs or benefits are immediate, people make poor choices, longrun payoffs being sacrificed in favor of immediate payoffs. For instance, smokers may know that their habit will harm them in the long run, but cannot bear to sacrifice the present urge to indulge in favor of the far-off reward of a healthy future. Similarly, a couple in the midst of sexual passion may appreciate that they may well regret their inadequate safety precautions at some point in the future, but they cannot control their present urges. We call this behavior *time-inconsistent*.[4]

[4] For an excellent survey of empirical results in this area, see Frederick, Loewenstein, and O'Donoghue (2002).

Are people time-consistent? Take, for instance, impulsive behavior. Economists are wont to argue that what appears to be impulsive—cigarette smoking, drug use, unsafe sex, overeating, dropping out of school, punching out your boss, and the like—may in fact be welfare-maximizing for people who have high time discount rates or who prefer acts that happen to have high future costs. Controlled experiments in the laboratory cast doubt on this explanation, indicating that people exhibit a *systematic* tendency to discount the near future at a higher rate than the distant future (Chung and Herrnstein 1967; Loewenstein and Prelec 1992; Herrnstein and Prelec 1992; Fehr and Zych 1994; Kirby and Herrnstein 1995; McClure et al. 2004).

For instance, consider the following experiment conducted by Ainslie and Haslam (1992). Subjects were offered a choice between \$10 on the day of the experiment or \$11 a week later. Many chose to take the \$10 without delay. However, when the same subjects were offered \$10 to be delivered a year from the day of the experiment or \$11 to be delivered a year and a week from the day of the experiment, many of those who could not wait a week *right now* for an extra 10%, preferred to wait a week for an extra 10%, provided the agreed-upon wait was one year in the future.

It is instructive to see exactly where the consistency conditions are violated in this example. Let x mean "\$10 at some time t" and let y mean "\$11 at time $t+7$," where time $t\psi$is measured in days. Then the present-oriented subjects display $x \succ y$ when $t\psi= 0$, and $y\psi\succ x$ when $t\psi= 365$. Thus the exclusion condition for $\succ$ is violated, and because the completeness condition for $\succeq$ implies the exclusion condition for $\succ$, the completeness condition must be violated as well.

However, time inconsistency *disappears* if we model the individuals as choosing over a slightly more complicated choice space in which the distance between the time of choice and the time of delivery of the object chosen is explicitly included in the object of choice. For instance, we may write x_0 to mean "\$10 delivered immediately" and x_{365} to mean "\$10 delivered a year from today," and similarly for y_7 and y_{372}. Then the observation that $x_0 \succ y_7$ and $y_{372} \succ x_{365}$ is no contradiction.

Of course, if you are not time-consistent and if you know this, you should not expect that your will carry out your plans for the future when the time comes. Thus, you may be willing to *precommit* yourself to making these future choices, even at a cost. For instance, if you are saving in year 1 for a purchase in year 3, but you know you will be tempted to spend the money in year 2, you can put it in a bank account that cannot be accessed until the year after next. My teacher Leo Hurwicz called this the "piggy bank effect."

The central theorem on choice over time is that time consistency results from assuming that *utility is additive across time periods and that the instantaneous utility function is the same in all time periods, with future utilities discounted to the present at a fixed rate* (Strotz 1955). This is called *exponential discounting* and is widely assumed in economic models. For instance, suppose an individual can choose between two consumption streams $x = x_0, x_1, \ldots$ or $y = y_0, y_1, \ldots$. According to exponential discounting, he has a utility function $u(x)$ and a constant $\delta \in (0,1)$ such that the total utility of stream x is given by[5]

$$U(x_0, x_1, \ldots) = \sum_{k=0}^{\infty} \delta^k u(x_k). \tag{1.1}$$

We call δ the individual's *discount factor*. Often we write $\delta = e^{-r}$ where we interpret $r > 0$ as the individual's one-period continuously compounded *interest rate*, in which case (1.1) becomes

$$U(x_0, x_1, \ldots) = \sum_{k=0}^{\infty} e^{-rk} u(x_k). \tag{1.2}$$

This form clarifies why we call this "exponential" discounting. The individual strictly prefers consumption stream x over stream y if and only if $U(x) > U(y)$. In the simple compounding case, where the interest accrues at the end of the period, we write $\delta = 1/(1+r)$, and (1.2) becomes

$$U(x_0, x_1, \ldots) = \sum_{k=0}^{\infty} \frac{u(x_k)}{(1+r)^k}. \tag{1.3}$$

Despite the elegance of exponential discounting, observed intertemporal choice for humans appears to fit more closely the model of *hyperbolic discounting* (Ainslie and Haslam 1992; Ainslie 1975; Laibson 1997), first observed by Richard Herrnstein in studying animal behavior (Herrnstein, Laibson, and Rachlin 1997) and reconfirmed many times since (Green et al. 2004). For instance, continuing the previous example, let z_t mean

[5] Throughout this text, we write $x \in (a,b)$ for $a < x < b$, $x \in [a,b)$ for $a \leq x < b$, $x \in (a,b]$ for $a < x \leq b$, and $x \in [a,b]$ for $a \leq x \leq b$.

"amount of money delivered $t\psi$days from today." Then let the utility of z_t be $u(z_t) = z/(t+1)$. The value of x_0 is thus $u(x_0) = u(10_0) = 10/1 = 10$, and the value of y_7 is $u(y_7)\psi= u(11_7)\psi= 11/8\psi= 1.375$, so $x_0 \succ y_7$. But $u(x_{365})\psi= 10/366\psi= 0.027\psi$while $u(y_{372})\psi= 11/373\psi= 0.029$, so $y_{372} \succ x_{365}$.

There is also evidence that people have different rates of discount for different types of outcomes (Loewenstein 1987; Loewenstein and Sicherman 1991). This would be irrational for outcomes that could be bought and sold in perfect markets, because all such outcomes should be discounted at the market interest rate in equilibrium. But, of course, there are many things that people care about that cannot be bought and sold in perfect markets.

Neurological research suggests that balancing current and future payoffs involves adjudication among structurally distinct and spatially separated modules that arose in different stages in the evolution of *H. sapiens* (Tooby and Cosmides 1992; Sloman 2002; McClure et al. 2004). The long-term decision-making capacity is localized in specific neural structures in the prefrontal lobes and functions improperly when these areas are damaged, despite the fact that subjects with such damage appear to be otherwise completely normal in brain functioning (Damasio 1994). *H. sapiens* may be structurally predisposed, in terms of brain architecture, to exhibit a systematic present orientation.

In sum, time inconsistency doubtless exists and is important in modeling human behavior, but this does not imply that people are irrational in the weak sense of preference consistency. Indeed, we can model the behavior of time-inconsistent rational individuals by assuming they maximize their time-dependent preference functions (O'Donoghue and Rabin, 1999a,b, 2000, 2001). For axiomatic treatment of time-dependent preferences, see Ahlbrecht and Weber (1995) and Ok and Masatlioglu (2003). In fact, humans are much closer to time consistency and have much longer time horizons than any other species, probably by several orders of magnitude (Stephens, McLinn, and Stevens 2002; Hammerstein 2003). We do not know why biological evolution so little values time consistency and long time horizons even in long-lived creatures.

1.5 Bayesian Rationality and Subjective Priors

Consider decisions in which a stochastic event determines the payoffs to the players. Let $X\psi$be a set of prizes. A *lottery* with payoffs in $X\psi$is a function $p\colon X \to [0,1]$ such that $\sum_{x\in X} p(x) = 1$. We interpret $p(x)$ as the probability that the payoff is $x\psi\in X$. If $X\psi= \{x_1,\dots,x_n\}$ for some finite number n, we write $p(x_i) = p_i$.

The *expected value* of a lottery is the sum of the payoffs, where each payoff is weighted by the probability that the payoff will occur. If the lottery $l\psi$has payoffs $x_1,\dots,x_n$ with probabilities $p_1,\dots,p_n$, then the expected value $\mathbf{E}[l]$ of the lottery $l\psi$is given by

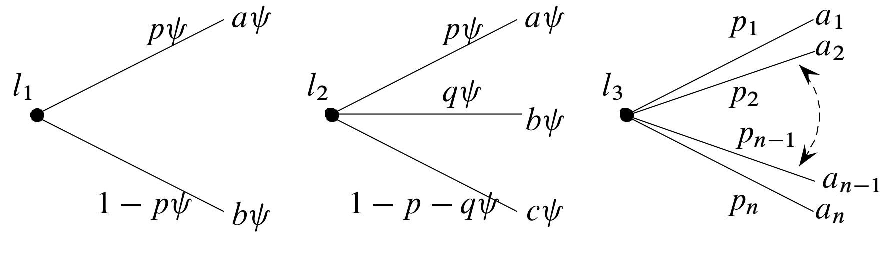
$$\mathbf{E}[l] = \sum_{i=1}^{n} p_i x_i.\psi$$

The expected value is important because of the law of large numbers (Feller 1950), which states that as the number of times a lottery is played goes to infinity, the average payoff converges to the expected value of the lottery with probability 1.

Consider the lottery l_1 in figure 1.1(a), where p is the probability of winning amount a and $1-p\psi$is the probability of winning amount b. The expected value of the lottery is then $\mathbf{E}[l_1]\psi= pa + (1-p)b$. Note that we model a lottery a lot like an extensive form game—except that there is only one player.

Consider the lottery l_2 with the three payoffs shown in figure1.1(b). Here $p\psi$is the probability of winning amount a, $q\psi$is the probability of winning amount b, and 1–p–q is the probability of winning amount c. The expected value of the lottery is $\mathbf{E}[l_2] = pa + qb + (1-p-q)c$.

A lottery with $n\psi$payoffs is given in figure 1.1(c). The prizes are now $a_1,\dots,a_n$ with probabilities $p_1,\dots,p_n$, respectively. The expected value of the lottery is now $\mathbf{E}[l_3] = p_1a_1 + p_2a_2 + \cdots + p_na_n$.



(a) (b) (c)

Figure 1.1. Lotteries with two, three, and n potential outcomes.

In this section we generalize the previous argument, developing a set of behavioral properties that yield both a utility function over outcomes and a probability distribution over states of nature, such that the expected utility principle holds. Von Neumann and Morgenstern (1944), Friedman and Savage (1948), Savage (1954), and Anscombe and Aumann (1963) showed that the expected utility principle can be derived from the assumption that individuals have consistent preferences over an appropriate set of lotteries. We outline here Savage's classic analysis of this problem.

For the rest of this section, we assume $\succeq$ is a preference relation (§1.1). To ensure that the analysis is not trivial, we also assume that $x\psi\succeq y$ is false for at least some $x, y\psi\in X$. Savage's accomplishment was to show that if the individual has a preference relation over *lotteries* that has some plausible properties, then not only can the individual's preferences be represented by a utility function, but also we can infer the probabilities the individual implicitly places on various events, and the expected utility principle holds for these probabilities.

Let Ω be a finite set of *states of nature*. We call $A\subseteq\Omega$ *events*. Let $\mathcal{L}$ be a set of lotteries, where a *lottery* is a function $\pi:\Omega\rightarrow X\psi$that associates with each state of nature $\omega\psi\in\Omega$ a payoff $\pi(\omega)\in X$. Note that this concept of a lottery does not include a probability distribution over the states of nature. Rather, the Savage axioms allow us to associate a subjective prior over each state of nature ω, expressing the decision maker's personal assessment of the probability that $\omega\psi$will occur. We suppose that the individual chooses among lotteries without knowing the state of nature, after which Nature chooses the state $\omega\psi\in\Omega$ that obtains, so that if the individual chose lottery $\pi\psi\in\mathcal{L}$, his payoff is $\pi(\omega)$.

Now suppose the individual has a preference relation $\succ$ over $\mathcal{L}$ (we use the same symbol $\succ$ for preferences over both outcomes and lotteries). We seek a set of plausible properties of $\succ$ over lotteries that together allow us to deduce (a) a utility function $u:X\rightarrow\mathbf{R}$ corresponding to the preference relation $\succ$ over outcomes in X; (b) a probability distribution $p:\Omega\rightarrow\mathbf{R}$ such that the expected utility principle holds with respect to the preference relation $\succ$ over lotteries and the utility function $u(\cdot)$; i.e., if we define

$$\mathbf{E}_\pi[u;p] = \sum_{\omega\in\Omega} p(\omega)u(\pi(\omega)),\psi \tag{1.4}$$

then for any $\pi,\rho\in\mathcal{L}$,

$$\pi\psi\succ\rho \Longleftrightarrow \mathbf{E}_\pi[u;p] > \mathbf{E}_\rho[u;p].\psi$$

Our first condition is that $\pi\psi\succ\rho$ depends only on states of nature where $\pi\psi$ and ρ have different outcomes. We state this more formally as follows.

A1. For any $\pi, \rho, \pi', \rho' \in \mathcal{L}$, let $A = \{\omega \in \Omega | \pi(\omega) \neq \rho(\omega)\}$. Suppose we also have $A = \{\omega \in \Omega | \pi'(\omega) \neq \rho'(\omega)\}$. Suppose also that $\pi(\omega) = \pi'(\omega)$ and $\rho(\omega) = \rho'(\omega)$ for $\omega \in A$. Then $\pi \succ \rho \Leftrightarrow \pi' \succ \rho'$.

This axiom says, reasonably enough, that the relative desirability of two lotteries does not depend on the payoffs where the two lotteries agree. The axiom allows us to define a *conditional preference* $\pi \succ_A \rho$, where $A \subseteq \Omega$, which we interpret as "π is strictly preferred to ρ, conditional on event A," as follows. We say $\pi \succ_A \rho$ if, for some $\pi', \rho' \in \mathcal{L}$, $\pi(\omega) = \pi'(\omega)$ and $\rho(\omega) = \rho'(\omega)$ for $\omega \in A$, $\pi'(\omega) = \rho'(\omega)$ for $\omega \notin A$, and $\pi' \succ \rho'$. Because of A1, this is well defined (i.e., $\pi \succ_A \rho$ does not depend on the particular $\pi', \rho' \in \mathcal{L}$). This allows us to define $\succeq_A$ and $\sim_A$ in a similar manner. We then define an event $A \subseteq \Omega$ to be *null* if $\pi \sim_A \rho$ for all $\pi, \rho \in \mathcal{L}$.

Our second condition is then the following, where we write $\pi = x|A$ to mean $\pi(\omega) = x$ for all $\omega \in A$ (i.e., $\pi = x|A$ means π is a lottery that pays x when A occurs).

A2. If $A \subseteq \Omega$ is not null, then for all $x, y \in X$, $\pi = x|A \succ_A \pi = y|A \Leftrightarrow x \succ y$.

This axiom says that a natural relationship between outcomes and lotteries holds: if π pays x given event A and ρ pays y given event A, and if $x \succ y$, then $\pi \succ_A \rho$, and conversely.

Our third condition asserts that the probability that a state of nature occurs is independent of the outcome one receives when the state occurs. The difficulty in stating this axiom is that the individual cannot choose probabilities but only lotteries. But, if the individual prefers x to y, and if $A, B \subseteq \Omega$ are events, then the individual treats A as more probable than B if and only if a lottery that pays x when A occurs and y when A does not occur is preferred to a lottery that pays x when B occurs and y when B does not. However, this must be true for any $x, y \in X$ such that $x \succ y$, or the individual's notion of probability is incoherent (i.e., it depends on what particular payoffs we are talking about—for instance, wishful thinking, where if the prize associated with an event increases, the individual thinks it is more likely to occur). More formally, we have the following, where we write $\pi = x, y|A$ to mean "$\pi(\omega) = x$ for $\omega \in A$ and $\pi(\omega) = y$ for $\omega \notin A$."

A3. Suppose $x\psi\succ y$, $x' \succ y'$, $\pi, \rho, \pi', \rho' \in \mathcal{L}$, and $A, B\psi\subseteq \Omega$. Suppose that $\pi\psi= x, y|A$, $\rho\psi= x', y'|A$, $\pi' = x, y|B$, and $\rho' = x', y'|B$. Then $\pi\psi\succ \pi' \Leftrightarrow \rho \succ \rho'$.

The fourth condition is a weak version of *first-order stochastic dominance*, which says that if one lottery has a higher payoff than another for any event, then the first is preferred to the second.

A4. For any event A, if $x\psi\succ \rho(\omega)$ for all $\omega\psi\in A$, then $\pi\psi= x|A \succ_A \rho$. Also, for any event A, if $\rho(\omega)\psi\succ x\psi$for all $\omega\psi\in A$, then $\rho \succ_A \pi\psi= x|A$.

In other words, if for any event A, $\pi\psi= x$ on A pays more than the best $\rho\psi$ can pay on A, then $\pi\psi\succ_A \rho$, and conversely.

Finally, we need a technical property to show that a preference relation can be represented by a utility function. We say nonempty sets $A_1, \ldots, A_n$ form a *partition* of set $X\psi$if the A_i are mutually disjoint ($A_i \cap A_j = \emptyset$ for $i\psi\neq j$) and their union is $X\psi$(i.e., $A_1 \cup \cdots\psi\cup A_n = X$). The technical condition says that for any $\pi, \rho\psi\in \mathcal{L}$, and any $x\psi\in X$, there is a partition $A_1, \ldots, A_n$ of Ω such that, for each A_i, if we change $\pi\psi$so that its payoff is $x\psi$on A_i, then $\pi\psi$is still preferred to ρ, and similarly, for each A_i, if we change $\rho\psi$so that its payoff is $x\psi$on A_i, then $\pi\psi$is still preferred to ρ. This means that no payoff is "supergood," so that no matter how unlikely an event A is, a lottery with that payoff when A occurs is always preferred to a lottery with a different payoff when A occurs. Similarly, no payoff can be "superbad." The condition is formally as follows.

A5. For all $\pi, \pi', \rho, \rho' \in \mathcal{L}$ with $\pi\psi\succ \rho$, and for all $x\psi\in X$, there are disjoint subsets $A_1, \ldots, A_n$ of Ω such that $\cup_i A_i = \Omega$ and for any A_i (a) if $\pi'\psi(\omega) = x$ for $\omega\psi\in A_i$ and $\pi'\psi(\omega) = \pi(\omega)$ for $\omega\psi\notin A_i$, then $\pi' \succ \rho$, and (b) if $\rho'(\omega)\psi= x\psi$for $\omega\psi\in A_i$ and $\rho'(\omega) = \rho(\omega)$ for $s\psi\notin A_i$, then $\pi\psi\succ \rho'$.

We then have Savage's theorem.

THEOREM 1.3 *Suppose A1–A5 hold. Then there is a probability function $p\psi$on Ω and a utility function $u : X \to \mathbf{R}$ such that for any $\pi, \rho\psi\in \mathcal{L}$, $\pi\psi\succ \rho\psi$ if and only if* $\mathbf{E}_\pi[u; p] > \mathbf{E}_\rho[u; p]$.

The proof of this theorem is somewhat tedious; it is sketched in Kreps 1988.

We call the probability $p\psi$the individual's *Bayesian prior*, or *subjective prior* and say that A1–A5 imply *Bayesian rationality*, because the they together imply Bayesian probability updating.

1.6 The Biological Basis for Expected Utility

Suppose an organism must choose from action set $X\psi$under certain conditions. There is always uncertainty as to the degree of success of the various options in X, which means essentially that each $x\psi\in X\psi$determines a lottery that pays $i\psi$offspring with probability $p_i(x)$ for $i\psi= 0,1,\ldots,n$. Then the expected number of offspring from this lottery is $(x)\psi= \sum_{j=1}^n jp_j(x)$. Let $L\psi$be a lottery on $X\psi$that delivers $x_i \in X\psi$with probability q_i for $i\psi= 1,\ldots,k$. The probability of $j\psi$offspring given L is then $\sum_{i=1}^k q_ip_j(x_i)$, so the expected number of offspring given L is

$$\sum_{j=1}^n j\sum_{i=1}^k q_ip_j(x_i) = \sum_{i=1}^k q_i\sum_{i=1}^k jp_j(x_i) = \sum_{i=1}^k q_i\quad (x_i),\psi \qquad (1.5)$$

which is the expected value theorem with utility function $(\cdot)$. See also Cooper (1987).

1.7 The Allais and Ellsberg Paradoxes

Although most decision theorists consider the expected utility principle acceptably accurate as a basis of modeling behavior, there are certainly well established situations in which individuals violate it. Machina (1987) reviews this body of evidence and presents models to deal with them. We sketch here the most famous of these anomalies, the *Allais paradox* and the *Ellsberg paradox*. They are, of course, not paradoxes at all but simply empirical regularities that do not fit the expected utility principle.

Maurice Allais (1953) offered the following scenario. There are two choice situations in a game with prizes $x\psi=$ \$2,500,000, $y\psi=$ \$500,000, and $z\psi=$ \$0. The first is a choice between lotteries $\pi\psi= y\psi$and $\pi' = 0.1x + 0.89y + 0.01z$. The second is a choice between $\rho = 0.11y + 0.89z\psi$ and $\rho' = 0.1x\psi+ 0.9z$. Most people, when faced with these two choice situations, choose $\pi\psi\succ \pi'$ and $\rho' \succ \rho$. Which would you choose?

This pair of choices is not consistent with the expected utility principle. To see this, let us write $u_h = u(2500000)$, $u_m = u(500000)$, and $u_l =$

$u(0)$. Then if the expected utility principle holds, $\pi\psi\succ \pi'$ implies $u_m >\psi$ $0.1u_h + 0.89u_m + 0.01u_l$, so $0.11u_m > 0.10u_h + 0.01u_l$, which implies (adding $0.89u_l$ to both sides) $0.11u_m + 0.89u_l > 0.10u_h + 0.9u_l$, which says $\rho \succ \rho'$.

Why do people make this mistake? Perhaps because of *regret*, which does not mesh well with the expected utility principle (Loomes 1988; Sugden 1993). If you choose π' in the first case and you end up getting nothing, you will feel really foolish, whereas in the second case you are probably going to get nothing anyway (not your fault), so increasing the chances of getting nothing a tiny bit (0.01) gives you a good chance (0.10) of winning the really big prize. Or perhaps because of *loss aversion* (§1.9), because in the first case, the anchor point (the most likely outcome) is \$500,000, while in the second case the anchor is \$0. Loss-averse individuals then shun π', which gives a positive probability of loss whereas in the second case, neither lottery involves a loss, from the standpoint of the most likely outcome.

The Allais paradox is an excellent illustration of problems that can arise when a lottery is consciously chosen by an act of will and one *knows* that one has made such a choice. The regret in the first case arises because if one chose the risky lottery and the payoff was zero, one knows for certain that one made a poor choice, at least ex post. In the second case, if one received a zero payoff, the odds are that it had nothing to do with one's choice. Hence, there is no regret in the second case. But in the real world, most of the lotteries we experience are chosen by default, not by acts of will. Thus, if the outcome of such a lottery is poor, we feel bad because of the poor outcome but not because we made a poor choice.

Another classic violation of the expected utility principle was suggested by Daniel Ellsberg (1961). Consider two urns. Urn A has 51 red balls and 49 white balls. Urn $B\psi$also has 100 red and white balls, but the fraction of red balls is unknown. One ball is chosen from each urn but remains hidden from sight. Subjects are asked to choose in two situations. First, a subject can choose the ball from urn A or urn B, and if the ball is red, the subject wins \$10. In the second situation, the subject can choose the ball from urn A or urn B, and if the ball is white, the subject wins \$10. Many subjects choose the ball from urn A in both cases. This violates the expected utility principle no matter what probability the subject places on the probability $p\psi$ that the ball from urn $B\psi$is white. For in the first situation, the payoff from choosing urn A is $0.51u(10)+0.49u(0)$ and the payoff from choosing urn $B\psi$

is $(1-p)u(10)+pu(0)$, so strictly preferring urn A means $p > 0.49$. In the second situation, the payoff from choosing urn A is $0.49u(10) + 0.51u(0)\psi$ and the payoff from choosing urn $B\psi$is $pu(10) + (1-p)u(0)$, so strictly preferring urn $A\psi$means $p\psi<\psi0.49$. This shows that the expected utility principle does not hold.

Whereas the other proposed anomalies of classical decision theory can be interpreted as the failure of linearity in probabilities, regret, loss aversion, and epistemological ambiguities, the Ellsberg paradox strikes even more deeply because it implies that humans systematically violate the following principle of first-order stochastic dominance (FOSD).

> Let $p(x)$ and $q(x)$ be the probabilities of winning x or more in lotteries A and B, respectively. If $p(x)\psi\succeq q(x)$ for all x, then $A \succeq B$.

The usual explanation of this behavior is that the subject *knows* the probabilities associated with the first urn, while the probabilities associated with the second urn are *unknown*, and hence there appears to be an added degree of risk associated with choosing from the second urn rather than the first. If decision makers are risk-averse and if they perceive that the second urn is considerably riskier than the first, they will prefer the first urn. Of course, with some relatively sophisticated probability theory, we are assured that there is in fact no such additional risk, it is hardly a failure of rationality for subjects to come to the opposite conclusion. The Ellsberg paradox is thus a case of performance error on the part of subjects rather than a failure of rationality.

1.8 Risk and the Shape of the Utility Function

If $\succeq$ is defined over X, we can say nothing about the *shape* of a utility function $u(\cdot)$ representing $\succeq$ because, by theorem 1.2, any increasing function of $u(\cdot)$ also represents $\succeq$. However, if $\succeq$ is represented by a utility function $u(x)$ satisfying the expected utility principle, then $u(\cdot)$ is determined up to an arbitrary constant and unit of measure.[6]

[6] Because of this theorem, the difference between two utilities means nothing. We thus say utilities over outcomes are *ordinal*, meaning we can say that one bundle is preferred to another, but we cannot say by how much. By contrast, the next theorem shows that utilities over lotteries are *cardinal*, in the sense that, up to an arbitrary constant and an arbitrary positive choice of units, utility is numerically uniquely defined.

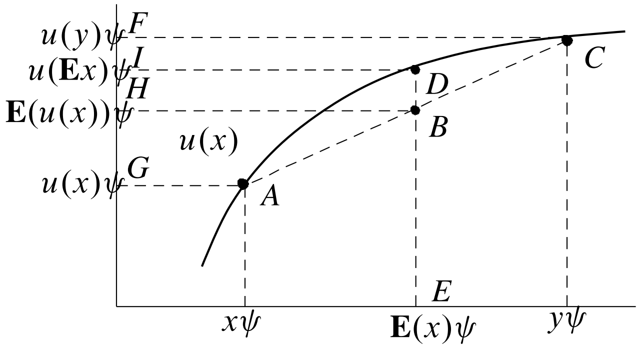


Figure 1.2. A concave utility function

Theorem 1.4 *Suppose the utility function $u(\cdot)$ represents the preference relation $\succeq$ and satisfies the expected utility principle. If $v(\cdot)$ is another utility function representing $\succeq$, then there are constants $a, b \in \mathbf{R}$ with $a > 0$ such that $v(x) = au(x) + b$ for all $x \in X$.*

For a proof of this theorem, see Mas-Collel, Whinston, and Green (1995, p. 173).

If $X = \mathbf{R}$, so the payoffs can be considered to be money, and utility satisfies the expected utility principle, what shape do such utility functions have? It would be nice if they were linear in money, in which case expected utility and expected value would be the same thing (why?). But generally utility is *strictly concave*, as illustrated in figure 1.2. We say a function $u : X \to \mathbf{R}$ is strictly concave if, for any $x, y \in X$ and any $p \in (0, 1)$, we have $pu(x) + (1 - p)u(y) < u(px + (1 - p)y)$. We say $u(x)$ is *weakly concave*, or simply *concave*, if $u(x)$ is either strictly concave or linear, in which case the above inequality is replaced by $pu(x) + (1 - p)u(y) = u(px + (1 - p)y)$.

If we define the lottery π as paying x with probability p and y with probability $1 - p$, then the condition for strict concavity says that *the expected utility of the lottery is less than the utility of the expected value of the lottery*, as depicted in figure 1.2. To see this, note that the expected value of the lottery is $E = px + (1 - p)y$, which divides the line segment between x and y into two segments, the segment xE having length $(px + (1 - p)y) - x = (1 - p)(y - x)$ and the segment Ey having length $y - (px + (1 - p)y) = p(y - x)$. Thus, E divides $[x, y]$ into two segments whose lengths have the ratio $(1 - p)/p$. From elementary geometry, it follows that B divides segment $[A, C]$ into two segments whose lengths have the same ratio. By the same reasoning, point H divides segments $[F, G]$ into segments with the same ratio of lengths. This means that point H has the coordinate value $pu(x) + (1-p)u(y)$, which is the expected utility of the lottery. But by definition, the utility of the expected value of the lottery is at D, which lies above H. This proves that the utility of the expected value is greater than the expected value of the lottery for a strictly concave utility function. This is know as *Jensen's inequality*.

What *are* good candidates for $u(x)$? It is easy to see that strict concavity means $u''(x) < 0$, providing $u(x)$ is twice differentiable (which we assume). But there are lots of functions with this property. According to the famous *Weber-Fechner law* of psychophysics, for a wide range of sensory stimuli and over a wide range of levels of stimulation, a just noticeable change in a stimulus is a constant fraction of the original stimulus. If this holds for money, then the utility function is logarithmic.

We say an individual is *risk-averse* if the individual prefers the expected value of a lottery to the lottery itself (provided, of course, the lottery does not offer a single payoff with probability 1, which we call a sure thing). We know, then, that an individual with utility function $u(\cdot)$ is risk-averse if and only if $u(\cdot)$ is concave.[7] Similarly, we say an individual is *risk-loving* if he prefers any lottery to the expected value of the lottery, and *risk-neutral* if he is indifferent between a lottery and its expected value. Clearly, an individual is risk-neutral if and only if he has linear utility.

Does there exist a measure of risk aversion that allows us to say when one individual is more risk-averse than another, or how an individual's risk aversion changes with changing wealth? We may define individual A to be *more risk-averse* than individual B if whenever A prefers a lottery to an amount of money x, B will also prefer the lottery to x. We say A is *strictly* more risk-averse than B if he is more risk-averse and there is some lottery that B prefers to an amount of money x but such that A prefers x to the lottery.

Clearly, the degree of risk aversion depends on the curvature of the utility function (by definition the *curvature* of $u(x)$ at x is $u''(x)$), but because $u(x)$ and $v(x) = au(x) + b$ $(a > 0)$ describe the same behavior, although $v(x)$ has curvature a times that of $u(x)$, we need something more sophis-

[7] One may ask why people play government-sponsored lotteries or spend money at gambling casinos if they are generally risk-averse. The most plausible explanation is that people enjoy the act of gambling. The same woman who will have insurance on her home and car, both of which presume risk aversion, will gamble small amounts of money for recreation. An excessive love for gambling, of course, leads an individual either to personal destruction or to wealth and fame (usually the former).

ticated. The obvious candidate is $\lambda_u(x) = -u''(x)/u'(x)$, which does not depend on scaling factors. This is called the *Arrow-Pratt coefficient of absolute risk aversion*, and it is exactly the measure that we need. We have the following theorem.

THEOREM 1.5 *An individual with utility function $u(x)$ is more risk-averse than an individual with utility function $v(x)$ if and only if $\lambda_u(x)\psi>\psi\lambda_v(x)\psi$ for all x.*

For example, the logarithmic utility function $u(x)\psi= \ln(x)\psi$has Arrow-Pratt measure $\lambda_u(x)\psi= 1/x$, which decreases with x; i.e., as the individual becomes wealthier, he becomes less risk-averse. Studies show that this property, called *decreasing absolute risk aversion*, holds rather widely (Rosenzweig and Wolpin 1993; Saha, Shumway, and Talpaz 1994; Nerlove and Soedjiana 1996). Another increasing concave function is $u(x)\psi= x^a$ for $a \in (0, 1)$, for which $\lambda_u(x) = (1-a)/x$, which also exhibits decreasing absolute risk aversion. Similarly, $u(x)\psi= 1 - x^{-a}$ $(a > 0)$ is increasing and concave, with $\lambda_u(x)\psi= -(a + 1)/x$, which again exhibits decreasing absolute risk aversion. This utility has the additional attractive property that *utility is bounded:* no matter how rich you are, $u(x) < 1$.[8] Yet another candidate for a utility function is $u(x) = 1-e^{-ax}$ for some $a > 0$. In this case $\lambda_u(x) = a$, which we call *constant absolute risk aversion.*

Another commonly used term is *coefficient of relative risk aversion*, $\mu_u(x) = \lambda_u(x)/x$. Note that for any of the utility functions $u(x) = \ln(x)$, $u(x) = x^a$ for $a \in (0, 1)$, and $u(x) = 1 - x^{-a}$ $(a > 0)$, $\mu_u(x)$ is constant, which we call *constant relative risk aversion*. For $u(x) = 1-e^{-ax}$ $(a > 0)$, we have $\mu_u(x) = a/x$, so we have *decreasing relative risk aversion*.

1.9 Prospect Theory

A large body of experimental evidence indicates that people value payoffs according to whether they are *gains* or *losses* compared to their current status quo position. This is related to the notion that individuals adjust to an accustomed level of income, so that subjective well-being is associated more with *changes* in income rather than with the *level* of income. See, for instance, Helson (1964), Easterlin (1974, 1995), Lane (1991, 1993), and

[8] If utility is unbounded, it is easy to show that there is a lottery that you would be willing to give all your wealth to play no matter how rich you are. This is not plausible behavior.

Oswald (1997). Indeed, people appear to be about twice as averse to taking losses as to enjoying an equal level of gains (Kahneman, Knetsch, and Thaler 1990; Tversky and Kahneman 1981b). This means, for instance, that an individual may attach zero value to a lottery that offers an equal chance of winning $1000 and losing $500. This also implies that people are *risk-loving over losses* while they remain risk-averse over gains (§1.8 explains the concept of risk aversion). For instance, many individuals choose a 25% probability of losing $2000 rather than a 50% chance of losing $1000 (both have the same expected value, of course, but the former is riskier).

More formally, suppose an individual has utility function $v(x-r)$, where $r\psi$is the status quo (his current position), and $x\psi$represents a change from the status quo. *Prospect theory*, developed by Daniel Kahneman and Amos Tversky, asserts that (a) there is a "kink" in $v(x-r)$ such that the slope of $v(\cdot)$ is two to three times as great just to the left of $x = r\psi$as to the right; (b) that the curvature of $v(\cdot)\psi$is positive for positive values and negative for negative values; and (c) the curvature goes to zero for large positive and negative values. In other words, individuals are two to three times more sensitive to small losses than they are to small gains, they exhibit declining marginal utility over gains and declining absolute marginal utility over losses, and they are very insensitive to change when all alternatives involve either large gains or large losses. This utility function is exhibited in figure 1.3.

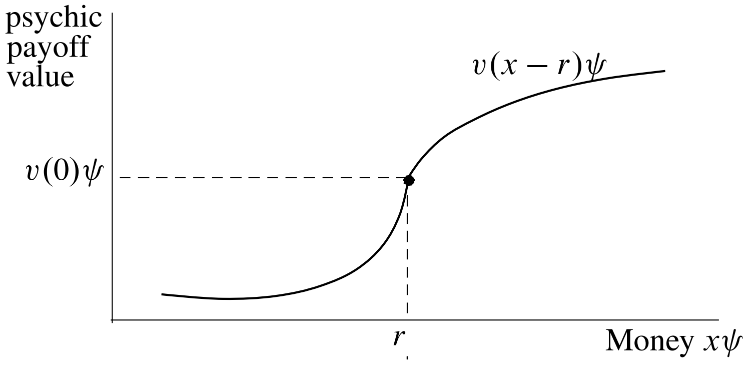


Figure 1.3. Loss aversion according to prospect theory

Experimental economists have long known that the degree of risk aversion exhibited in the laboratory over small gambles cannot be explained by standard expected utility theory, according to which risk aversion is measured by the curvature of the utility function (§1.8). The problem is that for small gambles the utility function should be almost flat. This issue has been formalized by Rabin (2000). Consider a lottery that imposes a $100 loss and offers a $125 gain with equal probability $p \not= 1/2$. Most subjects in the laboratory reject this lottery. Rabin shows that if this is true for all expected lifetime wealth levels less than $300,000, then in order to induce a subject to sustain a loss of $600 with probability 1/2, you would have to offer him a gain of at least $36,000,000,000 with probability 1/2. This is, of course, quite absurd.

There are many regularities in empirical data on human behavior that fit prospect theory very well (Kahneman and Tversky 2000). For instance, returns on stocks in the United States have exceeded the returns on bonds by about 8 percentage points, averaged over the past 100 years. Assuming investors are capable of correctly estimating the shape of the return schedule, if this were due to risk aversion alone, then the average individual would be indifferent between a sure $51,209 and a lottery that paid $50,000 with probability 1/2 and paid $100,000 with probability 1/2. It is, of course, quite implausible that more than a few individuals would be this risk-averse. However, a loss aversion coefficient (the ratio of the slope of the utility function over losses at the kink to the slope over gains) of 2.25 is sufficient to explain this phenomenon. This loss aversion coefficient is very plausible based on experiments.

In a similar vein, people tend to sell stocks when they are doing well but hold onto stocks when they are doing poorly. A kindred phenomenon holds for housing sales: homeowners are extremely averse to selling at a loss and sustain operating, tax, and mortgage costs for long periods of time in the hope of obtaining a favorable selling price.

One of the earliest examples of loss aversion is the *ratchet effect* discovered by James Duesenberry, who noticed that over the business cycle, when times are good, people spend all their additional income, but when times start to go bad, people incur debt rather than curb consumption. As a result, there is a tendency for the fraction of income saved to decline over time. For instance, in one study unionized teachers consumed more when next year's income was going to increase (through wage bargaining) but did not consume less when next year's income was going to decrease. We can explain this behavior with a simple loss aversion model. A teacher's utility can be written as $u(c_t - r_t) + s_t(1 + \rho)$, where c_t is consumption in period t, s_t is savings in period t, ρ is the rate of interest on savings, and r_t is the reference point (status quo point) in period t. This assumes that the marginal utility of savings is constant, which is a very good approximation. Now suppose the reference point changes as follows: $r_{t+1} = \alpha r_t + (1-\alpha)c_t$, where $\alpha \in [0,1]$ is an adjustment parameter ($\alpha = 1$ means no adjustment and $\alpha = 0$ means complete adjustment to last year's consumption). Note that when consumption in one period rises, the reference point in the next period rises, and conversely.

Now, dropping the time subscripts and assuming the individual has income M, so $c + s = M$, the individual chooses c to maximize

$$u(c-r) + (M - c)(1+\rho).$$

This gives the first order condition $u'(c-r) = 1+\rho$. Because this must hold for all r, we can differentiate totally with respect to r, getting

$$u''(c-r)\frac{dc}{dr} = u''(c-r).$$

This shows that $dc/dr = 1 > 0$, so when the individual's reference point rises, his consumption rises an equal amount.

One general implication of prospect theory is a *status quo bias*, according to which people often prefer the status quo over any of the alternatives but if one of the alternatives becomes the status quo, that too is preferred to any of the alternatives (Kahneman, Knetsch, and Thaler 1991). Status quo bias makes sense if we recognize that any change can involve a loss, and because on the average gains do not offset losses, it is possible that any one of a number of alternatives might be preferred if it is the status quo. For instance, if employers make joining a 401k savings plan the default position, almost all employees join. If *not* joining is made the default position, most employees do *not* join. Similarly, if the state automobile insurance commission declares one type of policy the default option and insurance companies ask individual policyholders how they would like to vary from the default, the policyholders tend not to vary, no matter what the default is (Camerer 2000).

Another implication of prospect theory is the *endowment effect* (Kahneman, Knetsch, and Thaler 1991), according to which people place a higher value on what they possess than they place on the same things when they do not possess them. For instance, if you win a bottle of wine that you could sell for $200, you may drink it rather than sell it, but you would never think of buying a $200 bottle of wine. A famous experimental result exhibiting the endowment effect was the "mug" experiment described by Kahneman, Knetsch and Thaler (1990). College student subjects given coffee mugs with the school logo on them demand a price two to three times as high to *sell* the mugs as those without mugs are willing to pay to *buy* the mugs. There is evidence that people underestimate the endowment effect and hence cannot appropriately *correct* for it in their choice behavior (Loewenstein and Adler 1995).

Yet another implication of prospect theory is the existence of a *framing effect*, whereby one form of a lottery is strictly preferred to another even though they have the same payoffs with the same probabilities (Tversky and Kahneman 1981a). For instance, people prefer a price of \$10 plus a \$1 discount to a price of \$8 plus a \$1 surcharge. Framing is, of course, closely associated with the endowment effect because framing usually involves privileging the initial state from which movements are assessed.

The framing effect can seriously distort effective decision making. In particular, when it is not clear what the appropriate reference point is, decision makers can exhibit serious inconsistencies in their choices. Kahneman and Tversky give a dramatic example from health care policy. Suppose we face a flu epidemic in which we expect 600 people to die if nothing is done. If program A is adopted, 200 people will be saved, while if program B is adopted, there is a 1/3 probability 600 will be saved and a 2/3 probability no one will be saved. In one experiment, 72% of a sample of respondents preferred A to B. Now suppose that if program C is adopted, 400 people will die, while if program D is adopted there is a 1/3 probability nobody will die and a 2/3 probability 600 people will die. Now, 78% of respondents preferred D to C, even though A and C are equivalent in terms of the probability of each final state, and B and D are similarly equivalent. However, in the choice between A and B, alternatives are over gains, whereas in the choice between C and D, the alternatives are over losses, and people are loss-averse. The inconsistency stems from the fact that there is no natural reference point for the decision maker, because the gains and losses are experienced by others, not by the decision maker himself.

The brilliant experiments by Kahneman, Tversky, and their coworkers clearly show that humans exhibit systematic biases in the way they make decisions. However, it should be clear that none of the above examples illustrates preference inconsistency once the appropriate parameter (current time, current position, status quo point) is admitted into the preference function. This point is formally demonstrated in Sugden (2003). Sugden considers a preference relation of the form $f\psi\succeq g|h$, which means "lottery $f\psi$is weakly preferred to lottery $g\psi$when one's status quo position is lottery h." Sugden shows that if several conditions on this preference relation, most of which are direct generalizations of the Savage conditions (§1.5), obtain, then there is a utility function $u(x,z)$ such that $f\psi\succeq g|h$ if and only if $\mathbf{E}[u(f,h)]\geq\mathbf{E}[u(g,h)]$, where the expectation is taken over the probability of events derived from the preference relation.

1.10 Heuristics and Biases in Decision Making

Laboratory testing of the standard economic model of choice under uncertainty was initiated by the psychologists Daniel Kahneman and Amos Tversky. In a famous article in the journal *Science*, Tversky and Kahneman (1974) summarized their early research as follows:

> How do people assess the probability of an uncertain event or the value of an uncertain quantity? ...people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors.

Subsequent research has strongly supported this assessment (Kahneman, Slovic, and Tversky 1982; Shafir and Tversky 1992; Shafir and Tversky 1995). Although we still do not have adequate models of these heuristics, we can make certain generalizations.

First, in judging whether an event A or object A belongs to a class or process B, one heuristic that people use is to consider whether A is *representative* of B but consider no other relevant facts, such as the frequency of B. For instance, if informed that an individual has a good sense of humor and likes to entertain friends and family, and asked if the individual is a professional comic or a clerical worker, people are more likely to say the former. This is despite the fact that a randomly chosen person is much more likely to be a clerical worker than a professional comic, and many people have a good sense of humor, so there are many more clerical workers satisfying the description than professional comics.

A particularly pointed example of this heuristic is the famous Linda the Bank Teller problem (Tversky and Kahneman 1983). Subjects are given the following description of a hypothetical person named Linda:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations.

The subjects were then asked to rank-order eight statements about Linda according to their probabilities. The statements included the following two:

> Linda is a bank teller.
> Linda is a bank teller and is active in the feminist movement.

More than 80% of the subjects—graduate and medical school students with statistical training and doctoral students in the decision science program at Stanford University's business school—ranked the second statement as more probable than the first. This seems like a simple logical error because every bank teller feminist is also a bank teller. It appears, once again, that subjects measure probability by *representativeness* and ignore baseline frequencies.

However, there is another interpretation according to which the subjects are correct in their judgments. Let p and q be properties that every member of a population either has or does not have. The standard definition of "the probability that member x is p" is the fraction of the population for which $p\psi$is true. But an equally reasonable definition is "the probability that $x\psi$is a member of a random sample of the subset of the population for which p is true." According to the standard definition, the probability of p and $q\psi$ cannot be greater than the probability of p. But, according to the second, the opposite inequality can hold: x might be more likely to appear in a random sample of individuals who are both $p\psi$and $q\psi$than in a random sample of the same size of individuals who are p. In other words, the probability that a randomly chosen bank teller is Linda is probably much lower than the probability that a randomly chosen feminist bank teller is Linda. Another way of expressing this point is that the probability that a randomly chosen member of the set "is a feminist bank teller" may be linda is greater than the probability that a randomly chosen member of the set "is a bank teller," is Linda.

A second heuristic is that in assessing the frequency of an event, people take excessive account of information that is easily *available* or highly *salient*, even though a selective bias is obviously involved. For this reason, people tend to overestimate the probability of rare events because such events are highly newsworthy while nonoccurrences are not reported. Thus, people worry much more about dying in an accident while flying than they do while driving, even though air travel is much safer than automobile travel.

A third heuristic in problem solving is to start from an initial guess, chosen for its representativeness or salience, and adjust upward or downward toward a final figure. This is called *anchoring* because there is a tendency to underadjust, so the result is too close to the initial guess. Probably as a result of anchoring, people tend to overestimate the probability of conjunctions (p and q) and underestimate the probability of disjunctions (p or q).

For an instance of the former, a person who knows an event occurs with 95% probability may overestimate the probability that the event occurs 10 times in a row, suggesting a probability of 90%. The actual probability is about 60%. In this case the individual starts with 95% and does not adjust downward sufficiently. Similarly, if a daily event has a failure one time in a thousand, people will underestimate the probability that a failure occurs at least once in a year, suggesting a figure of 5%. The actual probability is 30.5%. Again, the individual starts with 0.1% and doesn't adjust upward enough.

A fourth heuristic is that people prefer objective probability distributions to subjective distributions derived from applying probabilistic principles, such as the principle of insufficient reason, which says that if you are completely ignorant as to which of several outcomes will occur, you should treat them as equally probable. For example, if you give a subject a prize for drawing a red ball from an urn containing red and white balls, the subject will pay to have the urn contain 50% red balls rather than contain an indeterminate percentage of red balls. This is the famous Ellsberg paradox, analyzed in §1.7.

Choice theorists often express dismay over the failure of people to apply the laws of probability and conform to normative decision theory. Yet, people may be applying rules that serve them well in daily life. It takes many years of study to feel at home with the laws of probability, the understanding of which is the product of the last couple of hundred years of scientific research. Moreover, it is costly, in terms of time and effort, to apply these laws even if we know them. Of course, if the stakes are high enough, it is worthwhile to make the effort or engage an expert who will do it for you. But generally, as Kahneman and Tversky suggest, we apply a set of heuristics that more or less get the job done. Among the most prominent heuristics is simply *imitation*: decide what class of phenomenon is involved, find out what people normally do in that situation, and do it. If there is some mechanism leading to the survival and growth of relatively successful behaviors, and if the problem in question recurs with sufficient regularity, the choice-theoretic solution will describe the winner of a dynamic social process of trial, error, and imitation.

Should we expect people to conform to the axioms of choice theory—transitivity, independence from irrelevant alternatives, the sure-thing principle, and the like? Where we know that individuals are really optimizing, and have expertise in decision theory, we doubtless should. But this applies only to a highly restricted range of actions. In more general settings we should not. We might have recourse to Darwinian analysis, demonstrating that under the appropriate conditions individuals who are genetically constituted to obey the axioms of choice theory are better fit to solve general decision-theoretic problems and hence will emerge triumphant through an evolutionary dynamic. But human beings did not evolve facing general decision-theoretic problems. Rather, they faced a few specific decision-theoretic problems associated with survival in small social groups. We may have to settle for modeling these specific choice contexts to discover how our genetic constitution and cultural tools interact in determining choice under uncertainty.