3 Rational-Choice Theory

- **Overview**
- **Folk Psychology**
- **Ordinal-Choice Theory**
- **Cardinal-Choice Theory**
- **Stability, Invariance and Justifiers**
- **Conclusions**
- **Study Questions**
- **Suggested Readings**

Overview

According to one widely held view, human behavior is caused by intentions and motivations and therefore resists subsumption under natural laws. Rather, explanations of human behavior should describe the *reasons for actions*. The two chapters of this part deal with rationality at the individual level (Chapter 3) and at the level of interactions between individuals (Chapter 4). Both levels share the folk-psychological model according to which behavior is caused by the interaction of beliefs, opportunities and desires. They also share the idea that to explain an outcome means to rationalize it. According to this view, an account that portrays humans as erratic and unpredictable would not be explanatory as it would fail to elucidate the rational grounds for action.

Folk Psychology

There is a common-sense view that human behavior can and ought to be explained by the acting person's beliefs and desires. Why did Willy go on a diet? He believed he was overweight and hoped to lose some pounds. Why did Sally drink the water? Because she was thirsty and believed that

drinking water would quench her thirst. To give economic examples, why did the US Treasury allow the investment bank Lehman Brothers to fail? Because the then Treasurer, Hank Paulson, thought it would be bad to saddle taxpayers with paying to save a private company that screwed up, and he believed that bailing Lehman out would do just that. At a more theoretical level, why do businessmen invest in the production of a good up to a point where marginal costs equal marginal revenue? Because they aim to maximize profits.

The view that human behavior can and ought to be explained by citing beliefs and desires is called folk psychology. Sometimes folk psychology is understood as "the body of information people have about the mind," which in turn is regarded as the "basis for our capacity to attribute mental states and to predict and explain actions" (Nichols 2002: 134).

Although often treated as synonyms, it is useful to distinguish between mere behavior on the one hand and "action" or "choice" on the other. Behavior is the more general notion that describes physical movements of the body that originate within the individual. Not every movement of the body is also behavior. Traveling from Amsterdam to London by airplane moves the traveler's body but one wouldn't describe it as the traveler's behavior. Behavior is caused by the agent, as when I type these sentences or scratch my ear. Behavior can be intentional or unintentional. Twisting and turning during sleep or reflexes such as the sudden pulling away of one's hand from a hot stove constitute unintentional or "mere" behavior. Actions are intentional behavior, caused by the beliefs and desires of humans and thus the topic of folk psychology.

Decisions are somewhat in between beliefs and desires on the one hand and actions on the other. Sally's desire to drink some wine together with her belief that there is some chilled Chardonnay in the fridge explains her decision to get herself a glass. But decisions do not automatically lead to choices. She might make up her mind but then fail to convert the decision into action, because of weakness of will, forgetfulness or change of mind.

We have to impose some constraints on the beliefs and desires for them to serve as an explanation of the human action of interest. Sally's desire to be a superstar together with her belief that the Chinese have the best cuisine in the world doesn't explain her choice to go to Harvard rather than Columbia to study medicine. Beliefs and desires have to be connected with the chosen action in the right way in order to explain the action.

The typical form for a belief such that together with the agent's desire to *X* explains the action *A* is "*A* helps to promote *X*" or sometimes "*A* constitutes (the fulfillment of) *X*." Sally's wanting to be a superstar and her belief that performing with her band on TV will help her realize that goal jointly explain her decision to accept the offer from the TV station. In the case of choosing Harvard over Columbia the desire might be to go to the best medical school in the USA together with the belief that Harvard is the best school. In this case the action simply constitutes the fulfillment of her desire.

Beliefs and desires are thus *reasons for action*. But not every reason someone might have to act in certain ways also explains her actions. Sally might have perfectly good reasons to go to Harvard—her desire to go to the best medical school in the States, to become a highly acclaimed brain surgeon or what have you. But in fact she might decide to go to Harvard because she values the beauty of the Indian summers in Boston extremely highly. Thus, not every reason an individual might have to perform an action also constitutes the reason that explains his or her action. Rather, it is the reason the individual *acted on* that explains the action. When one acts on a reason such that the action is an effective means to one's end, one is said to act in accordance with instrumental rationality (Way forthcoming).

There are two ideas built into the concept of "acting on a reason" according to Donald Davidson: besides the idea of rationality there is also the idea of cause (Davidson 1974). He thinks that an explanatory reason is a rational cause. It is a cause in that it helps to *bring about* the action. It is neither necessary nor sufficient for the action. Had Sally not acted on her desire to admire Boston's Indian summers, she might have decided to go to Harvard anyway because it's the best school. And it's not sufficient because it will only jointly with many other factors determine the action (as mentioned, these factors will include the absence of weakness of will, for instance).

Rationality is required for a different reason. The motivations and beliefs that lead people to act are not typically transparent to outsiders such as social scientists. If Sally is our friend, we can ask her about her original motivation, and even if one cannot always take what people say at face value (and this is true even if we disregard intentional lying: people are often bad judges of their own motivations), we will have far more information that allows us to determine what was the actual reason she did act on than a social scientist analyzing an individual or a group or groups of individuals. A social scientist relies on evidence mostly in the form of observable behavior. But in order to infer motivations or beliefs from behavior (or other accessible forms of evidence), one must make fairly strong assumptions concerning the system of beliefs and desires people have. If individuals acted very erratically (though always on reasons!) it would be impossible to infer beliefs or desires or both from their actions (see Hausman 2000).

Models of rationality are essentially one way to constrain the beliefs and desires people are allowed to have in order for their actions to be explainable by a social scientist. In the next two sections I will describe in some detail two models of rationality that have received a great deal of attention in economics: a model of decision-making under certainty, *ordinal-choice theory*; and a model of decision-making under risk, *cardinal-choice theory*. In the following chapter I will go on to discuss decision-making in strategic situations, also known as *game theory*. First, though, let us look at decisions under certainty.

Ordinal-Choice Theory

Preferences

Economists explain action by preferences which represent beliefs and desires. Preferences differ in various respects from desires. Most fundamentally, preferences are comparative, desires are absolute. When one says that Sally desires to enjoy Indian summers, nothing is entailed about other desires she might hold. In particular, desires can conflict without contradiction: Sally might desire both the beauty of Boston Indian summers as well as the unique buzz of living in New York City, fully knowing that she can't have both, and by stating her desires she would not contradict herself. But she would contradict herself if she said she (strictly) prefers Boston to New York and New York to Boston.

What Sally might say without contradicting herself is that she prefers Boston to New York *qua* weather and New York to Boston *qua* buzz. We can call this concept of preferences "partial evaluative ranking." It ranks alternatives with respect to certain qualities or attributes people value. In this conception, people can have as many rankings as the alternatives have attributes people value.

People can also rank alternatives overall or "all things considered" (Hausman 2012). Apart from asking her which city she prefers *qua* weather and which *qua* buzz, we can ask her what she prefers all things considered. Ordinary language permits both uses of the term "prefers." Economists (and decision theorists) tend to employ the latter conception. For example, Richard Jeffrey, a well-known logician and decision theorist, wrote:

> But throughout, I am concerned with preference *all things considered*, so that one can prefer buying a Datsun to buying a Porsche even though one prefers the Porsche qua fast (e.g., since one prefers the Datsun qua cheap, and takes that desideratum to outweigh speed under the circumstances). *Pref* = preference *tout court* = preference on the balance.
>
> (Jeffrey 1990: 225; original emphasis)

The advantages of explaining action in terms of preferences rather than desires, and taking preferences to be "preferences on the balance" from the point of view of the economist or other social scientist are easy to see. Sally's desire to enjoy many Indian summers will not automatically explain her decision to move to Boston rather than New York because she might also have a desire to enjoy the New York City buzz. A preference for Boston over New York does not explain her decision, either, if it is a partial preference and therefore compatible with a partial preference the other way. But citing Sally's preference for Boston over New York *all things considered* goes a long way towards explaining her decision.

Thus far I have compared preferences with desires. A desire is a particular state of mind, a mental entity. Economists do not always feel comfortable when mental states are invoked in explaining phenomena of interest. Mental states are unobservable to everyone but the individual who has them, and therefore of dubious scientific value if one thinks that a proper science deals only with verifiable states of affairs, as, among others, the logical positivists did. Indeed, statements from papers written by the pioneers in what later has come to be known as the "revealed-preference theory" of consumer behavior give testimony that worries about introspection and the scientific status were among their motivations (all of the following quotations are taken from Sen 1973: 242). Paul Samuelson, for one, argued that he aimed to "develop the theory of consumer's behavior freed from any vestigial traces of the utility concept" (Samuelson 1938: 71). What he meant by "the utility concept" was the idea used by the classical utilitarians Jeremy Bentham, James and John Stuart Mill and Henry Sidgwick, and that was pleasure or happiness—a mental state (for more detailed discussions of utilitarianism, see Chapters 12 and 14). A decade later in a paper developing the revealed-preference theory Ian Little claimed "that a theory of consumer's demand can be based solely on consistent behavior," which for him meant that "the new formulation is scientifically more respectable [since] if an individual's behavior is consistent, then it must be possible to explain that behavior without reference to anything other than behavior" (I. Little 1949: 90, 97). Note the focus on explanation in this quotation. A final example is due to John Hicks, who said that: "the econometric theory of demand does study human beings, but only as entities having certain patterns of market behavior; it makes no claim, no pretence, to be able to see inside their heads" (Hicks 1956: 6).

In this early work on consumer demand, preferences were *identified* with choices. A clear statement of the identification comes from the paper by Ian Little already cited:

> The verb "to prefer" can either mean "to choose" or "to like better," and these two senses are frequently confused in the economic literature. The fact that an individual chooses *A* rather than *B* is far from conclusive evidence that he likes *A* better. But whether he likes *A* better or not should be completely irrelevant to the theory of price.
>
> (I. Little 1949: 91–2)

The idea that preference and choice are synonymous and that consumption theory can make do without non-choice data has become the "standard approach" to economic analysis. According to a recent paper:

> In the standard approach, the terms "utility maximization" and "choice" are synonymous. A utility function is always an ordinal index that describes how the individual ranks various outcomes and how he behaves (chooses) given his constraints (available options). The relevant

> data are revealed preference data; that is, consumption choices given the individual's constraints. These data are used to calibrate the model (i.e., to identify the particular parameters) and the resulting calibrated models are used to predict future choices and perhaps equilibrium variables such as prices. Hence, standard (positive) theory identifies choice parameters from past behavior and relates these parameters to future behavior and equilibrium variables.
>
> Standard economics focuses on revealed preference because economic data come in this form. Economic data can—at best—reveal what the agent wants (or has chosen) in a particular situation. Such data do not enable the economist to distinguish between what the agent intended to choose and what he ended up choosing; what he chose and what he ought to have chosen.
>
> (Gul and Pesendorfer 2008: 7–8)

Clearly, if economists want to explain economic phenomena using preferences they *must* be able to estimate preferences from data accessible to them, and individuals' mental states are not normally accessible. But the identification of preference with choice on which revealed-preference theory is based is too crude a means for achieving accessibility. Preferences are closely related to choices: preferences may *cause* and help to *explain* choices; preferences may be invoked to *justify* choices; in fortuitous circumstances, we can use preference data to make *predictions* about choices. But to identify the two would be a mistake (Hausman 2012: ch. 3).

To begin with, it is clear that we have preferences over vastly more states of affairs than we can ever hope (or dread) to be in the position to choose from. Here's a famous passage from Hume meant to illustrate a completely different point but which may serve as an example: "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (Hume 1960 [1739], "Of the Passions," part III, section 3). Most of us will have the opposite preference but never be in the position to choose between the two. In fact, most of the things we have preferences over we do not choose. I prefer to live in good health (rather than die a violent premature death), to have more money than Bill Gates (rather than have what I have), the next president of Russia to be sane (rather than insane). I can choose none of these things. I can choose the apple over the chocolate éclair for pudding; I can choose a career in the pharmaceutical industry over one in philosophy; I can choose to campaign for more democracy in Russia over staying put. But I never choose among those more ultimate things that concern me a great deal.

Economists may object that they are not in the business of providing an analysis of the ordinary concept of preference but rather in the business of science, where they can choose to define a technical concept as they please (as long as it is scientifically valuable). In other words, economists may *stipulate* a concept of preference which may only be loosely connected to our ordinary concept.

Unfortunately, the scientific value of the technical concept of preference as choice is dubious as well. One problem is that defining preference as choice makes it conceptually impossible for people to make counter-preferential choices (Sen 1977). And yet, counter-preferential choice is surely a genuine phenomenon. People make all sorts of mistakes when they choose due to inattentiveness, weakness of will or false beliefs. The other day I chose to watch the movie *J. Edgar* (directed by Clint Eastwood), believing that it would be a drama. I prefer drama to comedy (the only available alternative at the time). The movie turned out to be a romance, which I hate even more than comedies. I was also told that Leonardo DiCaprio had learned to act since *Titanic*. On top of being misled about the nature of the movie I was lied to. I counter-preferentially chose romance over comedy because I was ill-informed.

Economists could in principle stick to their guns and deny the existence (or economic importance) of counter-preferential choice. I hear there are still economists around who deny the existence (or economic importance) of involuntary unemployment or asset bubbles. That move would not help, however, because of an assumption economists make concerning preferences, and in fact need to make if they want to predict and explain choice behavior from observations of past choice behavior. The assumption is that preferences are *stable* during a reasonable passage of time. This is how Hal Varian puts it:

> When we talk of determining people's preferences from observing their behavior, we have to assume that the preferences will remain unchanged while we observe the behavior. Over very long time spans, this is not very reasonable. But for the monthly or quarterly time spans that economists usually deal with, it seems unlikely that a particular consumer's tastes would change radically. Thus we will adopt a maintained hypothesis that the consumer's preferences are stable over the time period for which we observe his or her choice behavior.
>
> (Varian 2010: 118)

If people are not allowed to make mistakes, it is very unlikely that preferences are stable, even over short periods of time. Three days ago I "preferred" comedy over romance, yesterday romance over comedy, and today comedy over romance again. Such "preferences" no one can work with. What happened in fact is that I have had a stable *mental ranking* of alternatives but made a mistake in my *choice* yesterday. A conception of preference as mental ranking is more useful (Hausman 2012).

Another reason why preferences shouldn't be understood as choices is that in the more interesting parts of economic theory, beliefs and expectations over future states of affairs are needed in addition to preferences in order to explain choices. This is certainly the case for decision-making under risk (see below) and for game theory (see Chapter 4). Beliefs and expectations

are mental states. To banish preferences understood as mental rankings because they are unobservable or subjective would mean one would have to banish beliefs and expectations too. One would throw out the baby with the bath water. Decisions under uncertainty and risk and game theory do not make sense without beliefs. And therefore preferences cannot be choices (cf. Hausman 2000, 2012). Henceforth, we will understand preferences as mental rankings of alternatives, "all things considered."

Choice Problems

The economic world does not come neatly parsed into clear-cut choice problems. Rather, an economist must formalize a given, naturally occurring situation into a choice problem. The way this is done determines the branch of choice theory relevant to the problem, and how the problem is to be solved. To give a simple and stupid example, suppose I am about to make breakfast and have to decide whether to have my coffee black (without milk) or white (with milk). One could formalize this as a simple choice between two goods (black coffee, white coffee) and apply decision theory under certainty. I might have a preference for white coffee over black, and since white coffee is available one could use decision theory to predict my choosing white coffee. But there are many more ways to conceive of the situation. We can build some uncertainty into it. I might prefer white coffee to black, but not when the milk is sour. I don't know for sure whether the milk is sour, but I can make reasonable guesses about it. If I know the probability that the milk is sour, then the problem is one of decision-making under risk, which will be examined in the next section.

There are yet other ways to think about the situation. Suppose I live with a flatmate, and he is responsible for buying milk. Now my best decision depends on his action, and his action might in turn depend on mine (he might or might not know for instance that I *really* like my coffee with milk and get very upset if there isn't any; his decision to buy milk might depend on his fear of my reproach). We're now in a game-theoretic situation.

Finally, my decision might depend on all sorts of contextual features. My decision to put milk in my coffee was immanent when I was making breakfast. That was an important piece of information, because I have white coffee for breakfast but black coffee after lunch. So my preferences are not over goods as such but rather over, say, "consumption bundles" which include relevant contextual features. An example due to Amartya Sen illustrates this point:

> Suppose the person faces a choice at a dinner table between having the last remaining apple in the fruit basket (y) and having nothing instead (x), forgoing the nice-looking apple. She decides to behave decently and picks nothing (x), rather than the one apple (y). If, instead, the basket had contained two apples, and she had encountered the choice between

> having nothing (*x*), having one nice apple (*y*) and having another nice one (*z*), she could reasonably enough choose one (*y*), without violating any rule of good behavior.
>
> (Sen 1993: 501)

The seemingly same act ("taking an apple from a fruit basket") can be a variety of different things, depending on whether there is more fruit in the basket but also of course on the social norms that are in place when the decision is being made—as Ken Binmore comments on Sen's example (Binmore 2009: 9): "The people in Sen's story inhabit some last bastion of civilization where Miss Manners still reigns supreme"—and other contextual features such as earlier decisions. Someone's preference for having an apple or not surely depends on whether he's already had a dozen or rather none, and if he is starving to death, he can safely decide to take the last apple even in the presence of Miss Manners. Care must be exercised when designing a choice problem.

For now I will ignore this issue and assume that the alternatives between which an agent chooses are sufficiently well described to apply decision theory coherently but come back to the issue further below.

Axioms and Preference Representation

Economists conceive of preferences as weak orders (in the mathematical or set-theoretic sense) over a set of available alternatives $x_1, x_2, \ldots, x_n$ in $\mathbf{X}$. I will use the symbol "≥" to mean "weakly prefers," that is, either "strictly prefers" or "is indifferent to." In order to constitute a weak order, preferences must satisfy a number of formal properties. One is transitivity:

> *Transitivity*: For all x_i, x_j, x_k in $\mathbf{X}$ if $x_i \geq x_j$, and $x_j \geq x_k$, then $x_i \geq x_k$.

If Sally prefers Harvard to Columbia, and Columbia to Johns Hopkins, she must also prefer Harvard to Johns Hopkins. The second main axiom is completeness:

> *Completeness*: For all x_i, x_j in $\mathbf{X}$, either $x_i \geq x_j$ or $x_j \geq x_i$ or both.

Completeness says that an agent is able to rank *all* available alternatives. For instance, Sally knows for any pair among the 134 institutions in the USA which award the degree of Doctor of Medicine whether she prefers one or the other or is indifferent between the two.

If one wants to represent preferences by means of a continuous utility function, as is often convenient, one has to assume that individuals' preferences satisfy an additional property:

> *Continuity*: For all x_j in $\mathbf{X}$, $\{x_i : x_i \geq x_j\}$ and $\{x_i : x_i \leq x_j\}$ are closed sets.

The axiom says that if an individual prefers each alternative in a series x_1, x_2, … to another alternative y, and the series converges to some alternative x_n, then the individual also prefers x_n to y.

When people's preferences satisfy these properties, they can be represented by a utility function that is unique up to a positive order-preserving transformation. What this means is that one can associate all available alternatives with numbers in such a way she strictly prefers an alternative with a higher number to an alternative with a lower number (and is indifferent between two alternatives with the same number). The association of numbers with alternatives is arbitrary as long as it preserves the order among the alternatives. Table 3.1 gives an example of an individual's preferences among brands of beer, where a brand that is higher up in the table (and is associated with a higher number) is preferred to any brand that appears lower in the table (and is associated with a lower number).

Table 3.1 Ordinal Utility

Brand	*Utility*		
Budvar	2	1,002	–11.8
Jupiler	1	1,001	–11.9
Carlsberg, Heineken	0	1,000	–12

This individual prefers Budvar to Jupiler and either beer to both Carlsberg and Heineken, and is indifferent between Carlsberg and Heineken. The different sets of numbers express nothing beyond this. In particular the absolute values of and differences or ratios between the levels of utility are meaningless. A number is only meaningful relative to the other numbers and only with respect to where it appears in the ranking. One can only address the question: "Is 1,002 more or less than or equal to 1,001?" not "How much more than 1,001 is 1,002?"

Transitivity and completeness are the main axioms of this model of choice. Are these axioms defensible? There are two main ways to defend them. We could either try to argue that the axioms are *normatively* accurate, in that a convincing case that people's preferences *ought to* satisfy them can be made. Or we could try to argue that the axioms are *descriptively* accurate, in that they are useful in predicting and explaining people's actual choices. Let us consider both kinds of defense.

Rationality and Ordinal-Choice Theory

The most common normative justification of the transitivity requirement is to point out that agents whose preferences are intransitive may be subject to exploitation. If Sally prefers Columbia to Johns Hopkins, she would probably

pay some money if she had a place at Johns Hopkins and was offered to swap. Now that she has a place at Columbia she'd pay some money for a place at Harvard. With intransitive preferences, she will now prefer Johns Hopkins to Harvard, once more pay money to get the place and end up where she started. This so-called "money-pump argument" in favor of transitivity was suggested by Frank Ramsey (1931 [1926]) and then developed by Davidson *et al.* (1955).

The money-pump argument is subject to a number of limitations, two of which I will consider here. First, people can be pumped only when they act on their preferences. Above I argued that preferences are not the same as choices. One might have intransitive preferences but never act on them and thus not be subject to exploitation. When one is offered trades, one might soon realize the risk of exploitation, amend one's preferences for the purpose of the trade and revert (or not) to intransitive preferences afterwards.

Second, the money-pump argument might be too strong to make its intended point. Let us suppose for the moment that an individual's preferences are indeed revealed by his choices. In order to prevent people from being money pumps, they do not necessarily have to have transitive preferences *at each point in time* but rather *over time*. Contrapositively, one can have transitive preferences at each point in time and still be victim to money pumpers because one's preferences change in such a way as to make them intransitive over time. Denote as "$>_t$" someone's preference at time t. Thus, an individual might have the following preferences: $x >_t y$, $y >_t z$ and $x >_t z$. That individual is in possession of z and is offered to trade it for y at an amount of money. He agrees. At time $t + 1$, the preferences have changed to: $x >_{t+1} z$, $z >_{t+1} y$ and $x >_{t+1} y$. He is offered a trade of the y that is now in his possession for x, and he agrees. At time $t + 2$, his preferences are now: $x >_{t+2} z$, $z >_{t+2} y$ and $x >_{t+2} z$. At this point, he is offered to trade the x that he now has for a z, which, once more, he agrees to. This individual's preferences are transitive throughout and yet he is being money-pumped because they are *dynamically* inconsistent. I will say a few more things about dynamic consistency below. For now, let me just state that there is nothing irrational as such with changing preferences.

Another argument that has been made is that the transitivity of preference is part of the meaning of the term "preference":

> The theory … is so powerful and simple, and so constitutive of concepts assumed by further satisfactory theory … that we must strain to fit our findings, or interpretations, to fit the theory. If length is not transitive, what does it mean to use a number of measure length at all? We could find or invent an answer, but unless or until we do, we must strive to interpret "longer than" so that it comes out transitive. Similarly for "preferred to."
>
> (Davidson 1980: 273; see also Broome 1991)

Davidson's defense is question-begging. If "preferred to" is analogous to "longer than," then "preferred to" must obey transitivity. But whether or not preference is relevantly like length is the question that is at stake here. We should not presuppose an answer.

Finally, there seem to be cases where decision-makers have good reason to entertain intransitive preferences. Paul Anand describes such a case:

> [I]magine that you are at a friend's dinner party and your host is about to offer you some fruit. If you are proffered an orange or small apple, you would rather have the orange, and if the choice is between a large apple and an orange you decide you would rather have the large apple. As it happens your friend is out of oranges and emerges from the kitchen with two apples, one large and one small. How should you choose? Etiquette seems to suggest that one might take the small apple and I find it difficult to see why such a choice must be judged irrational.
>
> (Anand 1993: 344)

The completeness property is even less well justifiable on rationality considerations (see for instance Elster 2007: 194; Gilboa *et al*. 2011). Robert Aumann once wrote:

> [O]f all the axioms of the utility theory, the completeness axiom is perhaps the most questionable. Like others, it is inaccurate as a description of real life; but unlike them we find it hard to accept even from the normative viewpoint.
>
> (Aumann 1962: 446)

If I'm offered "death by hanging" or "death by lethal injection" I might reasonably not have a preference for one over the other. And that wouldn't mean that I am *indifferent* between the two modes of dying. I am simply not able to rank the two options. Perhaps to the extent that preferences are used for explanations, this lack of justification does not matter too much. In a decision situation one is often forced to choose among alternatives, even in the absence of good reasons to go one way or the other. Perhaps economists are mainly interested in giving accounts for such situations. But, still, that is not a justification for the completeness axiom as an axiom of rationality.

There is an important difference between the absence of a preference between two options and indifference. Suppose one takes human life and money as incommensurable. One might then be given a choice between losing a human life and losing $10,000,000. One's absence of a preference then should not be interpreted as indifference, as the so-called "small-improvement argument" shows (Peterson 2009: 170). If one was really indifferent, then a small amount of money should tip the balance. So if one is really indifferent between saving a life and not expending $10,000,000,

then one should prefer not to expend $10,000,000 minus one cent. However, if one thinks of the two options as incommensurable, one will resist this conclusion.

In sum, the two major axioms of ordinal decision theory are not incontrovertible from the point of view of rationality. Meeting a money pumper and accepting his offers, one had better have dynamically consistent preferences. But this tells us little about other situations. Likewise, when one is forced to make a choice, one will, but the choice does not necessarily reveal a preexisting preference.

Ordinal-Choice Theory as Explanatory Theory

Actual decision-makers frequently violate the transitivity axiom. Economists have been concerned with the phenomenon since the late 1970s, when experimental work done by psychologists on the so-called "preference reversals" phenomenon was brought to their attention (Grether and Plott 1979). In the psychological experiments (Lichtenstein and Slovic 1971, 1973), subjects were asked to state their preference between lotteries. Pairs of lotteries were designed such that one offered a very high probability of winning a relatively small amount of money (the "P-bet"), and the other a smaller chance of winning a larger amount of money (the "$-bet"). The expected values of the two lotteries were roughly equal. Lichtenstein and Slovic predicted that people who chose P-bets would often pay more for a $-bet because in choice situations individuals' decisions are influenced primarily by the probability of winning or losing, whereas buying and selling prices are determined by dollar values. Their predictions were borne out in the data produced by their experiments.

Economists David Grether and Charles Plott began their paper with an acknowledgment of the significance of these findings:

> A body of data and theory has been developing within psychology which should be of interest to economists. Taken at face value the data are simply inconsistent with preference theory and have broad implications about research priorities within economics. The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind even the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior may result from principles which are of a completely different sort from those generally accepted.
>
> (Grether and Plott 1979: 623)

In their paper, Grether and Plott report the results of their own experiments in which they attempted to control for various alternative explanations of data such as insufficient incentives and indifference between lotteries (in Lichtenstein and Slovic's experiments, subjects did not have the option

of stating that they were indifferent between lotteries; Grether and Plott included that option but it was seldom taken). They concluded:

> Needless to say, the results we obtained were not those expected when we initiated this study. Our design controlled for all the economic-theoretic explanations of the phenomenon which we could find. The preference reversal phenomenon which is inconsistent with the traditional statement of preference theory remains.
>
> (Grether and Plott 1979: 634)

Economists take intransitive preferences seriously enough to develop choice theories without the transitivity axiom. One alternative to standard rational-choice theory that allows for intransitive preferences is regret theory (Loomes and Sugden 1982).

It is much harder to test the completeness axiom empirically because most economists take a very close relationship between choice and preference for granted. If, for instance, a subject refuses to choose between alternatives, this will often be interpreted as evidence for indifference. Nevertheless, Duncan Luce (2005 [1959]) observed that people sometimes seem to choose alternatives probabilistically rather than deterministically (e.g., x is chosen over y in p% of cases and y over x in $(1 - p)$% of cases). This could be interpreted as deterministic preferences switching back and forth all the time or, more plausibly, as stable *stochastic* preferences. Stochastic preferences conflict with the completeness axiom, which says that people always prefer either x over y or y over x or are indifferent between the two.

Cardinal-Choice Theory

The value of many of the consequences of our choices depends on factors we cannot influence and that we do not know with complete certainty. Suppose Marnix is planning a birthday party for his twins, and he has to choose whether to plan a trip to the municipal outdoor swimming pool or the bowling center. The twins, and thus Marnix, would strongly prefer going to the swimming pool, but only if the weather is sunny. If the weather is bad, this is the least preferred option. They rank bowling in between, and the "bad weather" alternative higher than the "good weather" alternative because of the regret they'd feel if they went bowling knowing how much they would have enjoyed swimming in the sun. How should they decide?

Risk and Uncertainty

Alas, no one knows for sure what the weather will be like. But we may reasonably assume that the uncertainty surrounding weather events can be described by a probability distribution over these events. In Frank Knight's terminology (Knight 1921), the decision-maker is thus facing *risk*, not

(radical) *uncertainty*. In decision-making under certainty, which outcome obtains is known. Sally, for example, was assumed to have full knowledge of the consequences of her choice between going to Harvard and going to Columbia. In decision-making under risk, it is not known what outcome will obtain, but it is known what outcome might obtain and with what probability. Outcomes are thus assumed to be generated by a stable process analogous to the rolling of a die or the spinning of a roulette wheel and the ball landing on a certain number. In decision-making under uncertainty it is neither known which outcome will obtain nor the probability with which it will occur. In fact, such a probability might not even exist. Arguably, most decisions actual economic agents face are characterized by this latter kind of uncertainty. In this book I will focus on decisions under risk because the associated decision theory is much better developed and easier to understand (but see Peterson 2009: ch. 3; Resnik 1987: ch. 2; Mitchell 2009).

Axioms and Preference Representation

There are three main differences between decision-making under certainty and under risk. First, the alternatives (over which the agent has preferences) are interpreted as *prospects*, which are defined as the pairing of the consequences of an action with the probabilities of these consequences occurring when the action is taken (Hargreaves Heap *et al*. 1992: 9). Essentially, prospects are lotteries. Suppose the probability of the weather's being good is p. The choice Marnix is facing is between one lottery that gives him p^*u(swimming | good weather) + $(1 - p)^*u$(swimming | bad weather), where $u(A \mid S)$ is the utility of action A given state of the world S, and another that gives him p^*u(bowling | good weather) + $(1 - p)^*u$(bowling | bad weather).

Second, in order to construct a representation of the agent's preferences by a(n expected) utility function, a variety of additional assumptions are required. The most important of these is an independence axiom, sometimes called Strong Independence (Hargreaves Heap *et al*. 1992: 9):

> *Strong Independence*: If $y = (x_i, x_j; p, 1 - p)$ and $x_i \sim y_i$, then $y \sim (y_i, x_j; p, 1 - p)$.

Strong Independence says that any component of a prospect can be replaced by another prospect to which the agent is indifferent, and the agent will be indifferent between the original and the new prospect. Hargreaves Heap *et al.* explain an implication of the axiom:

> Suppose that you are indifferent between \$100 for certain and a 50–50 chance of receiving \$250. Furthermore suppose that there are two prospects (I and II) which are identical except for one component: in I there is \$100 with probability 1/5 and in II there is a 50–50 chance of \$250

> with probability 1/5. Strong Independence implies that you will be indifferent between I and II, since they differ only with respect to this component and you are indifferent between the two options for this component. It is sometimes felt that this is an unreasonable inference, since the $100 is no longer certain in the comparison between I and II. Yet, is it really reasonable to have this indifference upset by the mere presence of other prizes? Strong Independence answers "no."
>
> (Hargreaves Heap *et al*. 1992: 10)

One of the reasons for mentioning Strong Independence here is that one of the most famous paradoxes in decision theory concerns a violation of the axiom (see below).

Third, if an agent's preferences satisfy all axioms, these can be represented by an expected-utility function that is *unique up to a positive affine transformation*. Thus, if an agent's preferences can be represented by an expected-utility function u, any function $u' = a + bu$ (where $a, b > 0$) can represent the agent's preferences equally well.

What an affine transformation means is best illustrated by an example of a quantity that is, like expected utility, measured on a cardinal scale: temperature. In order to determine a specific scale for measuring temperature, two points are fixed arbitrarily. In case of the Celsius scale, these are the melting point and the boiling point of water, and they are arbitrarily associated with 0° and 100°. Once these are fixed, however, any other temperature is determined. That the melting point of bismuth, for instance, is 271°C is not arbitrary, given the two fixed points of the Celsius temperature scale.

A couple of things are noteworthy about the cardinal scale temperature. First, as mentioned above, it is unique up to an affine transformation. For example, to convert Celsius (C) into Fahrenheit (F), the formula F = 32 + 9/5 C is used. Second, ratios of differences are meaningful. While it does not make sense to say either that it is twice as hot in New York (where temperature is measured in Fahrenheit) as in London (where it is measured in Celsius) nor that the difference between the temperature in New York and London is such-and-such, it is perfectly meaningful to say that the difference between London's temperature today and yesterday is twice as large as the difference between New York's temperature today and yesterday.

Using these properties of cardinal scales, one can construct a utility function from people's judgments and expressions of indifference as follows. Arbitrarily (but sensibly) fix the lowest- and highest-ranking alternatives as 0 and 1, respectively. In our example, u(swimming | good weather) = 1 and u(swimming | bad weather) = 0. Then ask Marnix, "At what probability of good weather p are you indifferent between going bowling for sure and playing a lottery of going swimming in good weather with probability p and going swimming in bad weather with probability $1 - p$?" Finally, define the expected utility of outcomes to be equal to that probability.

Risk Attitudes

Expected-utility functions have the so-called expected-utility property. That is, the utilities they assign to prospects are the sum of the utilities of the payoffs, weighted by their probabilities. Thus, if $w = [(x, p), (y, 1 - p)]$, then $EU(w) = p^*u(x) + (1 - p)^*u(y)$.

Using this property, we can define various attitudes towards risk, depending on how the expected utility of a prospect relates to the utility of its expected value. Three attitudes are usually distinguished:

- *Risk-neutrality* means that an agent is *indifferent* between playing a lottery and receiving the expected value of the prospect for certain; that is, his expected utility of the prospect is identical to the utility of its expected value: $EU(w) = p^*u(x) + (1 - p)^*u(y) = u(p^*x + (1 - p)y) = u(E(w))$. Firms in the theory of the firm (e.g., insurers) are often assumed to be risk-neutral.
- *Risk-aversion* means that an agent prefers receiving the expected value of the prospect for sure to playing the lottery: $EU(w)<u(E(w))$. Risk-aversion is often assumed for consumers. "Probabilistic insurances," for instance, in which the insured person receives the insurance sum with a probability less than 1 are rarely observed in the market. In the context of insuring their belongings, consumers are risk-averse.
- Being *risk-loving* means that an agent prefers playing the lottery to receiving its expected value for sure: $EU(w)>u(E(w))$. That some consumers are risk-loving must be assumed to explain gambling behavior because most gambles are "unfair" (gamblers receive less than the cost of playing the lottery on average). Interestingly, casino gambles are much fairer than the much more popular state lottery. In roulette, for example, the "house edge" (the average amount the player loses relative to any bet made) is a little above 5 percent in American roulette and 2.7 percent in European roulette, while the average player of a state lottery loses more than half of his ticket costs.

Expected-utility explanations are arguably somewhat deeper than explanations in terms of decision-making under certainty. To explain the choice of an apple if one could have had a banana by saying that the agent preferred an apple is not very illuminating. One way to interpret expected-utility theory (EUT) is to say that it constructs preferences over prospects from preferences over outcomes, given a risk attitude (cf. Hausman 2012). Assuming that people satisfy the axioms of EUT, they can be said to choose the prospect that maximizes their expected utility. But since the latter can be expressed as a weighted sum of utilities over outcomes, we can regard these utilities as basic and understand EUT as deriving preferences over prospects.

Consider a farmer who faces the choice between two crops, with the associated payoffs as described in Table 3.2.

Table 3.2 The Prudent Farmer

Weather	*Crop A (€; utility)*	*Crop B (€; utility)*
Bad (p = ½)	€10,000; 10	€15,000; 36
Good (1 – p = ½)	€30,000; 60	€20,000; 50
Average income	€20,000	€17,500
Average utility	35	43

We can easily see that the farmer is risk-averse because he derives higher utility from crop B, even though crop A gives him a higher average income.

To explain his choice, we can cite the preferences he has over the different outcomes and the beliefs he has about the probabilities of the weather. Most economists would say that the farmer's preferences over the prospects are given and basic. But this is implausible, and it prevents EUT being a genuinely explanatory theory. It is implausible because people will have more stable and basic preferences over things they ultimately care about. The farmer in this case cares about his income and the consumption associated with it, not about playing a lottery. (This may be different in other contexts. People might gamble solely for the enjoyment of the game and not for the money they might or might not win. In most cases, however, the enjoyment derives from the consequences of the choices, not from the choices themselves.)

The other reason for privileging this interpretation is explanation. If preferences over prospects were given, all an economist could say is that the farmer chose crop B because he preferred to do so—as in the case of decision-making under certainty. If one takes only preferences over outcomes as given and those over prospects or lotteries as derived, one can tell a more nuanced story about why farmer chose as he did.

Rationality and Expected-Utility Theory

The normative and descriptive aspects of expected-utility theory are interwoven, so I'll begin with a famous experimental observation of violation of Strong Independence: the Allais paradox (see Allais 1953). The Allais paradox is a choice problem designed by Nobel prize-winning economist Maurice Allais. Table 3.3 lists some of the choices subjects are given in experiments.

It turns out that most people choose A1 over A2, and most people choose A4 over A3. Importantly, the same individuals often choose A1 as well as A4, which violates Strong Independence. That axiom says that equal outcomes added to each of the two choices should have no effect on the relative desirability of one prospect over the other; equal outcomes should "cancel out." Experimental evidence suggests that they don't. People seem to prefer an amount for sure (€1,000 following act A1) to a gamble in which they have some chance of winning a higher amount but also some (albeit minimal)

Table 3.3 The Allais Paradox

States	*S1 (p = 0.89)*	*S2 (p = 0.1)*	*S3 (p = 0.01)*
Acts			
A1	€1,000	€1,000	€1,000
A2	€1,000	€5,000	€0
A3	€0	€1,000	€1,000
A4	€0	€5,000	€0

chance of winning nothing (A2). By contrast, if they are in a betting situation anyway (such as in the choice between A3 and A4), they prefer the lottery with the higher expected payoff.

It is important to see that it is not risk-aversion as such that explains these choices. Risk-aversion is consistent with expected-utility theory—the degree of risk-aversion can be measured by the curvature of the utility function (formally, it is measured by the ratio of its second derivative to its first derivative). Typical choices in the Allais paradox are *not* consistent with expected-utility theory. There is *no* utility function that is consistent with a preference of A1 over A2 and A4 over A3. To see this, compute the utility differences between the two pairs of acts (for any utility function):

$$u(\text{A1}) - u(\text{A2}) = u(€1\text{k}) - [0.89u(€1\text{k}) + 0.1u(€5\text{k}) + 0.01u(0)]$$
$$= 0.11u(€1\text{k}) - [0.1u(€5\text{k}) + 0.01u(0)]$$
$$u(\text{A3}) - u(\text{A4}) = 0.89u(0) + 0.11u(€1\text{k}) - [0.9u(0) + 0.1u(€5\text{k})]$$
$$= 0.11u(€1\text{k}) - [0.1u(€5\text{k}) + 0.01u(0)].$$

Many people confronted with these choices will, however, stick to them and insist that their choices are not irrational. Leonard Savage, one of the founders of modern decision theory, argued in response that in state S1 it does not matter, in either choice, which lottery is picked; this state should consequently be ignored. Decision-makers should base their decisions on features that differ between lotteries. In states S2 and S3 the payoffs differ between the lotteries, but the differences are exactly identical. Therefore, people should choose A1 over A2 if and only if they choose A3 over A4. This idea is called the "sure-thing principle" (Savage 1972: 21ff.).

Not everyone agrees with the sure-thing principle (McClennen 1988). In particular, it has been argued that Savage's principle begs the question as to why we ought to ignore the sure-thing outcomes. Perhaps Savage has given us an explanation of why violations sometimes occur, but he has not positively shown that we ought not to violate the principle. And there is certainly a relevant different between the pairs A1/A2 and A3/A4. If I were to end up in state S3 after choosing A2, I will regret my choice a great deal. I could

have had a good amount of money for sure. I chose to gamble and lost. That was silly. In the choice between A3 and A4, the odds that I end up with nothing are overwhelming anyway. I'd consider myself lucky if I did win but not winning wasn't silly. Quite to the contrary, it would have been unreasonable to forfeit a good chance of a considerable higher gain for a minimally smaller chance of losing. I would not regret my choice.

Expected-Utility Theory as Explanatory Theory

There are various other paradoxes like Allais', one of which I will discuss here. I will only point out that people often violate the axioms of expected-utility theory, but not ask whether it is reasonable to do so.

Ellsberg's Paradox. Ellsberg's paradox (see Ellsberg 1961), first noticed by Daniel Ellsberg when he was a PhD student in Harvard in the 1950s, also demonstrates a violation of Strong Independence. It involves choosing from an urn with different-colored balls whose composition is not precisely known. It therefore involves uncertainty, and not mere risk.

In the example, you are supposed to have an urn containing 30 red balls and 60 other balls that are either black or yellow. You know that there are 60 black and yellow balls in total, but not how many of each there are. The urn is well mixed, so each individual ball is as likely to be drawn as any other. You are now given two choices between two prospects each:

> Choice 1
> Option A: You receive €100 if you draw a red ball.
> Option B: You receive €100 if you draw a black ball.
>
> Choice 2
> Option C: You receive €100 if you draw a red or yellow ball.
> Option D: You receive €100 if you draw a black or yellow ball.

Since the prizes are exactly the same, it follows from EUT that you will prefer prospect A to prospect B if and only if you believe that drawing a red ball is more likely than drawing a black ball. Further, there would be no clear preference between the choices if you thought that a red ball was as likely as a black ball. Similarly it follows that you will prefer prospect C to prospect D if and only if you believe that drawing a red or yellow ball is more likely than drawing a black or yellow ball. It might seem intuitive that, if drawing a red ball is more likely than drawing a black ball, then drawing a red or yellow ball is also more likely than drawing a black or yellow ball. So, supposing you prefer prospect A to prospect B, it follows that you will also prefer prospect C to prospect D. In experiments, however, most people strictly prefer prospect A to prospect B and prospect D to prospect C.

To see that this violates EUT, again compute the differences of expected utilities between the lotteries:

$u(\text{A}) - u(\text{B}) = \frac{1}{3}u(€100) - p_{\text{Black}}u(€100)$
$u(\text{C}) - u(\text{D}) = \frac{1}{3}u(€100) + (\frac{2}{3}-p_{\text{Black}})u(€100) - \frac{2}{3}u(€100) = \frac{1}{3}u(€100) - p_{\text{Black}}u(€100)$.

People choose in such a way as to avoid gambles with unknown probabilities. Since the proportion of black balls is not known, they choose A over B. Since the proportion of yellow balls is not known, they choose D over C. This too violates the sure-thing principle.

Stability, Invariance and Justifiers

Whether or not to accept experimental data as violations of expected-utility theory depends on the experimenter's beliefs about the stability of subjects' preferences and about how subjects construe a choice problem (see the section on "Choice Problems" above). The two assumptions interact. Any *apparent* violation of an axiom of the theory can always be interpreted as any of three things:

- the subjects' preferences *genuinely* violate the axioms of the theory;
- the subjects' preferences have changed during the course of the experiment;
- the experimenter has overlooked a relevant feature of the context that affects the subjects' preferences.

As we saw above, economists assume that subjects' preferences are stable for the goals and purposes of an economic investigation. Let us formulate this idea as a principle (cf. Binmore 2009: 9):

> *Stability*. Individuals' preferences are stable over the period of the investigation.

Stability is not enough, however. This is because even assuming stability any apparent violation could be explained by the fact that a subject interprets the choice situation differently than the experimenter; she sees a difference between two choices where the experimenter sees none. Thus let us formulate a second principle (cf. Hausman 2012: 16):

> *Invariance*. Individuals' preferences are invariant to irrelevant changes in the context of making the decision.

I have given some examples of contextual features that do not appear to be irrelevant to subjects' preference orderings above. But not just any contextual feature is allowed to change the preference ordering. Economists, for instance, insist that the presenting of *irrelevant alternatives* should not matter to one's preferences, as illustrated by an anecdote involving the late Sidney

Morgenbesser, a philosopher at Columbia University, who, apparently, is better remembered for his wit than his publications, and about whom the following tale is told:

> According to the story, Morgenbesser was in a New York diner ordering dessert. The waitress told him he had two choices, apple pie and blueberry pie. "Apple," Morgenbesser said.
>
> A few minutes later the waitress came back and told him, oh yes, they also have cherry pie.
>
> "In that case," said Morgenbesser, "I'll have the blueberry."
>
> (Poundstone 2008: 50)

In the story, the availability of a further alternative, cherry pie, should not matter to the preference between apple and blueberry pie. Such a preference change, induced by an irrelevant alternative becoming available, is regarded as irrational by economists.

What features shall we allow to induce preference changes? It is clearly *not* irrational to prefer to drive on the right side of the road when on the Continent and on the left when in Britain. Nor is it irrational to prefer coffee to tea for breakfast and tea to coffee at teatime. It is also not irrational to prefer having a piece of chocolate cake to having nothing when that piece is the first and to prefer having nothing to having another piece when one has already had four. The mere passage of time does *not* seem to be a choice-relevant factor, however (as is illustrated by the fact that most economists regard hyperbolic discounting as an anomaly; see Chapter 15).

These considerations also show that there is a problem, which one could formulate as a dilemma for the economist. Standard rational-choice theory is usually regarded as a formal, as opposed to substantive, theory of rationality (e.g., Hausman and McPherson 2006; Hausman 2012). Here is one way of putting the issue:

> [T]hat an agent is rational from [rational-choice theory]'s point of view does not mean that the course of action she will choose is objectively optimal. Desires do not have to align with any objective measure of "goodness": I may want to risk swimming in a crocodile-infested lake; I may desire to smoke or drink even though I know it harms me. Optimality is determined by the agent's desires, not the converse.
>
> (Paternotte 2011: 307–8)

The idea goes a long way back, to both David Hume and Max Weber. Hume thought that "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume 1960 [1739]). People value this and that; reason has no say in what they ought to value. I should mention in passing that an assumption along the lines that people always prefer more money to less would be inconsistent with this

Humean principle. If economists have nothing to say about what individuals ought to value, they surely cannot assume that individuals always value more money higher than less.

Max Weber's influence stems from his view of objectivity in the social sciences. Weber thought, like David Hume, that there was a clear distinction between facts and values (Weber 1949). The social sciences, being sciences of human behavior, cannot avoid dealing with values altogether. Weber then thought that the social sciences can preserve objectivity by restricting the values that are allowed to play a role in scientific investigations to the values held by agents under investigation. The social scientist should not influence the study by adding his or her own values. Rather, he or she should take the analyzed agents' ends as given and proceed from there.

Against the backdrop of Hume's ideas about "reason versus passions" and Weber's views on objectivity, we can easily see the significance of the distinction between formal and substantive theories of rationality. Rationality is clearly an evaluative notion. A rational action is one that is commendable, and an irrational action is one that is not. One cannot consistently say that a certain choice would be irrational and at the same time that the agent ought to do it. But, according to the economist's view, it is the agent's values that matter in the evaluation, not the economist's. The economist provides only some formal constraints of consistency.

The problem is that invariance is not a merely formal principle. If we left it to the agent to determine what counts as a "relevant" feature of the context, no choice would ever be irrational. Preferring beer to wine at one instant and wine to beer at the next will not reveal intransitive preferences, because the agent will be a few heartbeats older, and he might consider that fact relevant (age is surely relevant in the limit: there is no inconsistency in preferring sweet to savory as a child and savory to sweet as an adult).

To see how difficult it can be to determine whether a contextual feature is relevant or not, consider an example that is very similar to Morgenbesser's choice between apple and blueberry pie. Recall that Morgenbesser was making fun of someone violating invariance by reversing his preference when a new option became available. The example considered now shows that it doesn't always seem irrational to reverse one's preferences when new options become available or unavailable. The next day, the waitress asks Morgenbesser if he'd like chicken or steak. He chooses steak. After a minute, the waitress comes back and says they have a daily special, USDA prime rib. Morgenbesser says he would like that instead. Another minute passes and the waitress comes back again, announcing that the last prime rib has just gone to the customer who is sitting at the table next to his. "In which case," Morgenbesser says, "I'd like to have the chicken."

Now that he could have had prime rib, every bite of the ordinary steak would remind him of the forgone opportunity and make him feel regretful. He therefore chooses chicken in order to avoid such feelings of regret. It is at least not clear that this should be regarded as a piece of flawed reasoning.

Hence, sometimes the becoming available or unavailable of an alternative can induce a rational preference change, at other times a change in preference is irrational.

This is not merely a philosopher's worry. In Chapter 10 below I will describe in more detail a series of experiments on intransitive preferences conducted by economists Graham Loomes, Chris Starmer and Robert Sugden (Loomes *et al*. 1991). Loomes *et al*. point out that the results of earlier experiments on intransitive preferences such as the preference-reversal experiments by Slovic and Lichtenstein or Grether and Plott can be explained away by accounts other than ones involving intransitive preferences. One of the alternative accounts is that people regard choice-tasks and valuation-tasks as different kinds of problems, and consequently have different preferences. Another is that subjects do not regard the series of tasks as independent but instead treat it as a single lottery. In both cases (and others) the transitivity axiom can be saved. (Albeit at the expense of violating Strong Independence; that axiom is, however, more controversial anyway.) In their own experiments, Loomes *et al*. control for these and other alternative explanations of the results. But they too must make assumptions such as stability and invariance for the preferred interpretation of their results because there are always some differences between any two choice situations.

John Broome has a principle, similar to invariance, that helps to the construction of choice problems (Broome 1991: 103):

> *Principle of Individuation by Justifiers*. Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them.

Broome's principle makes plain that one needs to make assumption about the nature of rationality when one designs a choice problem. Economists will not like this principle because they do not like to make substantive assumptions about rationality, which is why I concealed that matter by using a seemingly more innocuous term such as "relevant." But relevance, too, is something the economist has to decide on the basis of considerations about the nature of rationality.

The dilemma the economist faces, then, is this. He can either stick with the "formal axioms" of completeness, transitivity, Strong Independence and so on and refuse to assume the principles of stability and invariance. But then rational-choice theory will be useless for all explanatory and predictive purposes because people could have fully rational preferences that constantly change or are immensely context-dependent. Alternatively he can assume stability and invariance but only at the expense of making rational-choice theory a substantive theory, a theory laden not just with values but with *the economist's* values. The economist then has to decide whether, say, presenting an analogous problem as a choice-task and as a valuation-task is the same thing; more generally, whether framing a problem one way or another may reasonably affect someone's preferences; what relevant alternatives are;

whether, to what extent and what social norms may matter; whether to conceive of a series of choices as a single-choice problem or indeed a series of independent choices; and so on.

Conclusions

The empirical violations of rational-choice theory give the whole idea of explanation by reasons a somewhat ironic twist. The reason to look for rational-choice explanations of actions is that, at least according to Donald Davidson, there are no strict laws that relate mental events such as beliefs and desires with physical events such as bodily behavior. But the project of trying to explain human behavior is not yet doomed because we can explain behavior by citing reasons for action.

Social scientists must learn the reasons for action from observable behavior or otherwise accessible evidence. This means that they have to impose fairly stringent consistency and stability constraints on behavior in order for it to be interpretable in terms of rational-choice theory. But of course, human behavior isn't particularly consistent and stable—this is why there are no psycho-physical laws to begin with. If human behavior is not consistent and stable, the project of trying to explain it in terms of reasons for action is somewhat less promising.

Study Questions

1 Ought your preferences to be complete? Ought they to be transitive? Discuss.
2 Do we normally know the probability of outcomes which obtain in the future? How good is expected-utility theory as a model of choice?
3 Explain the difference between the Allais and Ellsberg paradoxes.
4 What are, in your view, contextual features that should make a difference to an individual's preference ordering?
5 Are unstable preferences irrational?

Suggested Readings

The best introduction to rational-choice theory is, to my mind, Resnik 1987. Somewhat more advanced and critical is Hargreaves Heap *et al*. 1992. Other very useful texts include Gilboa 2010 and Peterson 2009. Hausman 2012 provides a comprehensive treatment of the nature of preferences and the use of the concept in economics. Ratcliffe 2007 is a good and critical discussion of folk psychology.