1 Rationality, Preferences, and Utility Theory

Mainstream economics portrays individual agents as choosing rationally. Many of its generalizations concerning how people actually choose are also claims about how agents *ought rationally* to choose. This fact distinguishes economics from the natural sciences, whose particles do not choose and are neither rational nor irrational, and whose theories have no similar normative aspect. Chemists offer no advice to benzine molecules, which would not listen to advice if given. I have a good deal to say in Chapters 4, 13, and 16 about the significance of this distinctive feature of economics. In this chapter, my goal is to describe the fundamental elements of models of both rational and actual choice. Most of the chapter is devoted to the simplest model: "ordinal utility theory." However, Section 1.3 provides a sketch of expected utility theory, which is central to decision theory and plays an important role in mainstream economics.

1.1 Rational Choice with Perfect Knowledge: Preferences and Ordinal Utility Theory

What is it to choose rationally? This is an old philosophical question, which, like other old philosophical questions, is hard to answer. One can say, accurately, albeit unhelpfully, that rational choice consists in *choice that is properly responsive to reasons*. There are many ways to fail to be properly responsive to reasons and thus many kinds of irrationality. Furthermore, the notion of choice is ambiguous. It can refer to deliberating, or it can refer to the action that is the outcome of deliberation. Economists regard choice as action and regard it as determined by three factors: physical constraints, beliefs

https://doi.org/10.1017/9781009320283.003 Published online by Cambridge University Press

(or expectations), and preferences. Choices are rational if they are governed by rational preferences and rational beliefs. Noneconomists take "preferences" to be subjective states of individuals, which are reflected in their words and actions. Although preferences in economics differ from preferences in ordinary discourse in ways to be explained later, this chapter argues that preferences in economics, like preferences in ordinary discourse, are *subjective states that combine with beliefs to cause choices.*

If people are approximately rational, then a model of rational choice can be used to predict actual choice. A normative theory is concerned with value – that is, with what is good or bad – and with which actions are obligatory, permissible, or impermissible. Unlike "positive" theories that describe, predict, and explain what actually happens, normative theories evaluate what happens and say what ought to happen. Rationality is a normative notion, although not a moral notion. To fail to do what one rationally ought to do is foolish or self-defeating rather than evil.

This sketch of the distinction between positive and normative inquiries is subject to caveats. Among other difficulties, there is no sharp boundary between positive and normative. Just consider statements such as "members of the SS were cruel" or "Margaret Thatcher was shrewd." They state matters of fact, but they also offer evaluations. However, for our purposes, the rough distinction between what is, on the one hand, normative, prescriptive, or evaluative and, on the other hand, positive or factual will serve. As we shall see, the normative model of rational preference, belief, and choice this chapter presents can also play a central role in positive economics when joined with the hypothesis that people are largely rational.

The objects of choice can be many different things. In consumer choice theory, they are limited to bundles of commodities and services. In the theory of the firm, the alternatives may be combinations of inputs. Preferences range more widely. An individual, Marty, may have preferred that Hillary Clinton be elected president in 2008, that

Apple stock double in value, or that no hurricane strikes Puerto Rico, but none is a state of affairs that Marty can choose.

The description of the objects of both choice and preference must include "everything that matters to the agent" (Arrow 1970, p. 45). Otherwise, preferences would change with context. For example, I have no preference among the alternatives described merely as "a cup of coffee" or "a bottle of beer." Which I prefer depends on the time of day, what I am eating, what the weather is like, and many other things. The states of affairs ranked by preferences must instead be described as "drinking a cup of coffee versus a bottle of beer at 7 a.m. with cereal and …" or "drinking a bottle of beer versus a cup of coffee on a hot afternoon after mowing the lawn and …." I often simplify and speak of a preference for beer rather than speaking of a preference for the complete state of the world with drinking a beer versus the complete state of the world without doing so.[1]

The economist's model of rational choice largely abstracts from deliberation: constraints and beliefs fix which alternative actions are feasible and believed to be feasible, and agents choose whatever action is at the top of their already given preference ranking of the actions they believe to be feasible. Taken by itself, Yolanda's preference for blueberries over strawberries is not subject to rational appraisal, but there are rational constraints on sets of preferences. For example, if she also prefers cherries to blueberries, then she ought to prefer cherries to strawberries. The model of rational choice does not condemn as irrational Peter's preference for a side serving of mouse droppings over a portion of carrots, but it does find it irrational if Peter also prefers being healthy to being unhealthy.

[1] Johanna Thoma (2021b) argues that decision theory is not continuous with everyday explanations of behavior, on the grounds that the objects of preference in decision theory are context independent and hence maximally finely individuated. In my view, Thoma is making too much of an idealization.

1.1.1 *Certainty and Perfect Knowledge*

In unusual circumstances in which agents possess complete knowledge and there is neither risk nor uncertainty, what agents believe coincides with the facts, and nothing need be said about belief, rational or otherwise. The account of rationality in these circumstances is called "ordinal utility theory." Economists have a simple model of rational choice shown in Figure 1.1. Agents who have complete knowledge rank the alternatives among which they choose (represented here by different foods). Constraints may rule out some alternatives (bread in this case). Agents choose from the remaining options whatever is at the top of their preference ranking. In positive economics, an agent's preference ranking governs the agent's choices. In normative (welfare) economics, the objective is to help people move up their preference ranking. The principles of positive microeconomics are mainly generalizations concerning preferences and their implications for choice. The imperatives of normative economics specify how best to satisfy preferences. Preferences lie at the core of mainstream economics.

1.1.2 *Preference Axioms*

Mainstream economists agree on the following axioms concerning preferences in the special circumstances in which there is no uncertainty and agents possess perfect knowledge. Because, as Section 1.1.3 explains, preferences that satisfy these axioms can be represented by an ordinal utility function, these are called the axioms of ordinal utility theory. Although economists agree on these axioms, few think these axioms are universal truths. Some

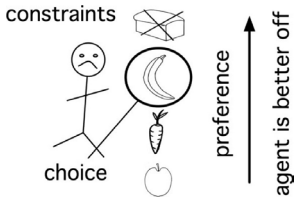

FIGURE 1.1 Preference and choice.

https://doi.org/10.1017/9781009320283.003 Published online by Cambridge University Press

economists believe that the axioms of ordinal utility theory are good approximations and that the violations can be regarded as unsystematic noise. Many question whether these axioms are generally true of people and regard them more as a point of comparison than as a guide to reality. Even those who have the fewest qualms about the model recognize that these axioms are simplifications of a more complicated reality. This is not just an armchair observation. As discussed in Chapters 13 and 14, there are experimental data revealing systematic violations of these axioms, and psychologists and behavioral economists have formulated generalizations concerning preferences that explain these violations. Nevertheless, for most economists, even behavioral economists, these axioms are the standard starting place for theorizing concerning individual choice.

The following two axioms (quoted from Mas-Colell et al. 1995, p. 6) are ubiquitous:

(*Completeness*) For all x, y in X, either $x \succeq y$ or $y \succeq x$ or both.

(*Transitivity*) For all x, y, and z in X if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.

"X" is the set of alternatives over which agents have preferences – commodity bundles in the case of consumer choice theory – and x, y, and z are alternatives in X. According to Mas-Colell et al., "[w]e read $x \succeq y$ as 'x is at least as good as y'" (1995, p. 6; see also Varian 1984, p. 111). This definition of "$x \succeq y$" might seem surprising, since the axioms are supposed to govern *preferences* within the (positive) science of economics, not judgments of goodness. It is better to read "$x \succeq y$" as "the agent either prefers x to y or is indifferent between x and y." "$x \succ y$" means "the agent prefers x to y," and "$x \sim y$" means that the agent is indifferent between x and y. Employing the weak preference relation "$\succeq$" is convenient, because one does not have to specify separately the transitivity of strong preference, indifference, and mixtures of the two, such as the claim that if $x \succ y$ and $y \sim z$, then $x \succ z$.

Varian (1984, pp. 111–12) includes two additional axioms, which, as I explain shortly, are needed to prove a crucial theorem:

> (*Reflexivity*) For all x in X, $x \succeq x$.
> (*Continuity*) For all y in X, $\{x : x \succeq y\}$ and $\{x : x \preceq y\}$ are closed sets.[2]

Reflexivity is trivial and arguably a consequence of completeness, while continuity is automatically satisfied for any finite set of alternatives.

In contrast to Varian, who presents the axioms as assumptions about people's actual preferences, Mas-Colell et al. maintain that completeness and transitivity are axioms of rationality: people's preferences are *rational* if they satisfy the axioms (1995, p. 6). Since Mas-Colell and his co-authors are concerned to offer an account of people's actual preferences, they must also maintain that to some extent people's preferences are in this sense rational.

1.1.3 Utilities and the Ordinal Representation Theorem

The ordinal representation theorem proves that when people's preferences satisfy these axioms,[3] then they can be represented by a continuous utility function that is unique up to a positive monotone (order-preserving) transformation (Debreu 1959, pp. 56–7). The "utility" of an alternative merely indicates the alternative's place

[2] A set is closed if it includes its boundaries. See Debreu 1959, pp. 54–9 and Harsanyi 1977b, p. 31. Suppose that cars varied continuously in both their fuel efficiency and their acceleration (which allows for an uncountable infinity of cars). If Helen has lexicographic preferences among cars – in particular if she ranked cars exclusively by their fuel efficiency and then by their acceleration only as a tie-breaker – she would violate the continuity axiom. If one were to draw a graph with fuel efficiency on the horizontal axis and acceleration on the vertical axis, with the point $(x^\star, y^\star)$ marking the efficiency and acceleration of a particular car, the vertical line $x = x^\star$ marks the boundary between the set of all acceleration–efficiency pairs Helen weakly prefers to $(x^\star, y^\star)$ and the set of pairs to which Helen prefers $(x^\star, y^\star)$. $x = x^\star$ belongs to neither set. Thus Helen violates the continuity axiom.

[3] The version of the theorem, proven by Debreu (1959, pp. 56–7) employs the additional technical condition that the set of bundles of the k commodities be a connected subset of R^k (the k-dimensional space of real numbers). A subset of R^k is "connected" if it is not the union of two nonempty disjoint and closed subsets of R^k.

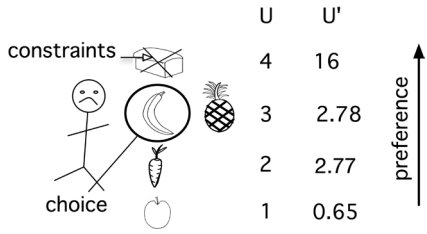

FIGURE 1.2 Ordinal utility.

in an agent's preference ranking. It is not something people seek or accumulate.

Here is a simple way to understand how a utility function "represents" preferences and what it means for it to be unique up to a positive order-preserving transformation. Suppose that an agent, Jill, who has preferences over a finite set of alternatives, adopts the convention of listing the alternatives on lined paper with preferred alternatives in higher rows and alternatives among which she is indifferent in the same row. Since Jill's preferences are complete, every alternative must find a place in the list. Since Jill's preferences are transitive, no alternative can appear in more than one row. Given such a list, one can assign numbers arbitrarily to rows, with the proviso that higher numbers are assigned to higher rows. Any numbering of the rows that is consistent with the ordering is an ordinal utility function. The numbers – the utilities – merely indicate where alternatives are located in Jill's preference ranking. Utility is not pleasure or usefulness or anything substantive at all. It is merely an indicator of an alternative's location in a preference ranking. Figure 1.2 provides an illustration of how ordinal utilities represent preferences.

The pictures of food represent the ordered list of alternatives. U and U' are two of the infinite number of utility functions that assign higher numbers to alternatives in higher rows, and the same number to alternatives in the same row. The numbers are arbitrary apart from their order. In Figure 1.2, Jill chooses the banana rather than an apple because she prefers it to the apple. The picture says nothing about why she prefers the banana to the apple; it certainly does not say that the reason is that the banana has more utility. That claim mistakenly

supposes that utility is something like pleasure, which is found in different quantities in the objects of preference. Utility is an indicator of preference. It is not an object of preference.

Jill does not choose the bread, despite preferring the bread to the banana, because she cannot have the bread, perhaps because the store has run out of bread or because she cannot afford to purchase it. Because she is indifferent between the banana and the pineapple, she could just as well have chosen the pineapple.

1.1.4 Further Assumptions Concerning Preferences

Economists make other assumptions governing preferences in addition to the axioms listed earlier. Some of these are occasionally called "axioms," but most often these assumptions about preferences are implicit. Here is a list:

1. Preferences are stable and "given" – that is, known and fixed before individuals choose. Preferences may change, but only infrequently. Because economists take preferences as given, it appears that they have nothing to say about how preferences are formed or modified. However, it is also the case that preferences among the immediate objects of choice depend on beliefs about their consequences and preferences among their consequences.
2. Preferences are independent of context or framing; they depend exclusively on the alternative states of affairs to be ranked.
3. Preferences are independent of irrelevant alternatives. If an agent prefers x from the set of alternatives $\{x, y\}$, then the agent does not prefer y from a larger set of alternatives including x and y, and if an agent prefers x from any set of alternatives including x and y, then the agent does not prefer y from the set $\{x, y\}$.
4. Preferences determine choices: among the alternatives they believe to be accessible, agents choose one that is at the top of their preference ranking.[4] This assumption, which I call "choice dependence," provides the crucial link between preference and choice.

[4] Mas-Collel et al. never state such an axiom explicitly. Varian expresses it informally as "[o]ur basic hypothesis is that a rational consumer will always choose a most preferred bundle from the set of feasible alternatives" (Varian 1984, p. 115).

Identifying these additional assumptions concerning preferences helps to pin down the concept of preferences that economists rely on. This is true even though these further assumptions, like the axioms, are problematic. Experiments carried out by psychologists and behavioral economics cast doubt on these further claims about preferences, especially the first two. The first assumption reveals an internal conflict. If economists can link preferences among the immediate objects of choice to preferences among their consequences and beliefs about the probabilities of those consequences, then they have something to say about preference formation and modification, and preferences are not merely given.

It is fortunate that economists have something to say about the formation and revision of preferences. If economists had nothing to say about what determines preferences among the immediate objects of choice, then their explanations and predictions would be trivial. In every case, the explanation for why an agent chose action A would be "the agent preferred A to the alternatives." To explain or to predict any choice would be merely to point to its location atop the ranking of feasible alternatives. There would be nothing to say about what determines and changes the preference ranking. For example, economists would be unable to predict how preferences among investors in a company's stock change with the settlement of a lawsuit against the company.

The second assumption of context independence is vulnerable to experimental critiques, and it is scarcely tenable even as an extreme idealization. This unavoidable complication risks trivializing conditions on rational choice. Suppose, for example, that Jack has intransitive preferences. He prefers x to y, y to z, and z to x. However, if "x when the alternative is y" and "x when the alternative is z" are different states of affairs, x_1 and x_2 respectively, then Jack prefers x_1 to y, y to z, and z to x_2, and the violation of transitivity has disappeared. To block this trivialization requires a substantive principle requiring indifference between alternatives such as x_1 and x_2. John Broome (1991b, pp. 103–4) argues for "a rational

requirement of indifference" such as "[o]utcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them" (1991b, p. 103). Whether it is rational to have a preference between two outcomes depends on a substantive theory of rationality.

Choice determination is of special importance. On the assumption of complete knowledge, there is no need to mention beliefs. But restating choice determination more simply as "agents choose an alternative at the top of their ranking of feasible alternatives" contributes to the mistaken espousal of revealed preference theory, which is discussed in Section 1.2. Agents can prefer x to y, yet choose y from the set of alternatives $\{x, y\}$, because they falsely believe themselves to be choosing from some other set of alternatives such as $\{z, y\}$.

What I have called "choice determination" is often called "utility maximization." Choosing an alternative that is at the top of one's preference ranking among feasible alternatives is choosing to maximize utility, but the terminology can be misleading. When economists say that individuals maximize utility, they are only saying that people do not rank any feasible option above the option they choose. Although the "utility" language was inherited from the utilitarians, some of whom thought of utility as a sensation with a certain intensity, duration, purity, or propinquity (Bentham 1789, chapter 4), there is no such implication in contemporary microeconomic theory. Economists sometimes speak misleadingly of individuals as seeking more utility, but they do not mean that utility is an object of choice: some ultimately good thing that people want in addition to good health or a faster internet connection. The theory of rational choice specifies no distinctive aims that all people must embrace. Utility is just an indicator of where an alternative is located within a preference ranking. Individuals who are utility maximizers just do what they most prefer. To say that individuals are utility maximizers says nothing about the nature of their preferences. *All it does is connect preference and choice (or action) in a particularly simple*

way. Rational individuals rank available alternatives and *choose* what they most *prefer* from among the alternatives they believe to be feasible.

1.1.5 Ordinal Utility Theory as a Theory of Rationality?

Because rationality is a normative notion, ordinal utility theory, as a theory of rational choice, is a normative theory. It purportedly tells us what our preferences should be like and how they should influence our choices. To define what rational preference and choice are is ipso facto to say how one ought rationally to prefer and to choose.

With the additional claim that people are in fact (approximately) rational in the sense just defined, utility theory implies a positive theory concerning how constraints, choice, preference, and belief are related. Utility theory, as a positive theory of preference and choice, is a crucial part of consumer choice theory. Because most of the axioms and the additional assumptions of utility theory appear to be false, there are many questions to ask about the role of ordinal utility theory in the explanation and prediction of economic phenomena. Part III addresses these methodological questions and considers the significance of the two faces of ordinal utility theory as both a theory of actual and of rational choice. Let us ask here merely whether the model of choice presented by ordinal utility theory is a plausible normative theory of *rational* choice. Is it irrational to violate its axioms and implicit conditions?

Some of the elements of ordinal utility theory are not intended as substantive principles of rationality. They function instead to define and simplify the domain to which the theory applies. For example, agents who deliberate about their preferences rather than taking them as given are not behaving irrationally. Requiring that preferences be already fixed is instead intended to separate the questions of interest to economists from other questions about decision-making. Although rationality may require some stability in preferences, there is nothing irrational in changing one's preferences. The assumption of stability serves mainly to make the theory usable and to limit the

circumstances to which the theory applies. Similarly, there seems to be nothing irrational in the inability to rank some alternatives, which violates completeness. However, one can regard completeness as a boundary condition on rational choice. If people cannot compare alternatives, then they cannot choose on the basis of reasons. Similarly, it is hard to see what would be irrational about violating continuity (Elster 1983, p. 8). But rather than regarding continuity as a boundary condition, one can regard it as trivial, because it is automatically satisfied if the set of alternatives is not uncountably infinite. Choice determination is questionable, too, but one can regard it more as a modeling decision than as a substantive requirement. By taking preferences to encompass everything that influences choices other than beliefs and constraints, only random errors fail to satisfy it.

One can make a plausible case that the remaining conditions are requirements of rationality. Reflexivity only demands indifference between identical alternatives. If preferences (as I argue) constitute or imply judgments about which alternatives are better, then, as John Broome argues (1991a), transitivity is implied by the logic of comparative adjectives such as "better than," and transitivity is hence a demand of rationality. Nevertheless, it would be surprising if experimenters could not find intransitivities in everybody's preferences among a sufficiently long and complicated series of choices among pairs of options. But, like miscalculations in arithmetic, the mistakes people make in following rules do not show that the rules themselves are mistaken. In defense of transitivity, one can also argue that, if our preferences fail to be transitive, then others can make fools of us. Suppose, for example, that I prefer x to y and y to z and z to x, and that I start out possessing z. Then I should, in principle, be willing to pay a fee for each of the following three exchanges: trade z away for y, trade y away for x, and trade x away for z. I am then back where I started, except that I am poorer by the amount of the expense of the three fees. I have become a "money pump," and this argument is known as the money pump argument. (See Schick 1986 for a critical discussion.) Transitivity appears to be a requirement of rationality.

If one relaxes the simplifications, takes a step toward greater realism, and recognizes that people typically do not have a ready-made preference ordering to guide their choice, then, as Herbert Simon argues (1982), it may be rational to adopt strategies that reduce the cognitive burden of decision-making and take account of the limits to one's information and information-processing abilities. Adopting these strategies will sometimes lead people to choose options that are later ascertained to be inferior to feasible alternatives. To economize on deliberation and to be a predictable partner in collective enterprises, it may also be rational to carry through with one's intentions or plans, even if changing course appears to be more advantageous. However, if one happens to have a preference ranking handy that actually manages to satisfy all the conditions concerning preferences and choices, then it is rational to allow one's preferences to determine one's choices.

These comments explain why economists regard ordinal utility theory as a fragment of a theory of rational choice that specifies conditions that preferences must satisfy in order to justify choices. This theory of rational choice purports to be purely *formal* and to say nothing about what things it is rational to prefer. Because it is purely formal, this view of rationality might be regarded as too weak. As just noted, without substantive assumptions that rule out some preferences as irrational, the axioms turn out to be trivial. And it seems that some preferences, such as Derek Parfit's example of "future Tuesday indifference" (1984, p. 124 – indifference to anything that happens on a future Tuesday), should be regarded as irrational, regardless of their consistency with other preferences.

Critics have also argued that this model of rational choice is too demanding. Must an agent A be able to rank all feasible options, or is it enough that A be able to rank all the options that are available in the given context or in some set of alternatives worth considering? Is full transitivity necessary or is it enough that A's choices never form a cycle? Such possible weakenings of the standard axioms have their own formal developments, and one can prove a variety of

theorems relating these conceptions to each other (see Sen 1971 and McClennen 1990, chapter 2). Most economic theory relies on standard ordinal utility theory, and the details of formal developments of weaker alternatives are not germane here.

1.2 REVEALED PREFERENCE THEORY

Revealed preference theory is an interpretation of formal results explored initially by Paul Samuelson (1938, 1947), generalized and developed by many others (especially Houthakker 1950), and elegantly summarized by Arrow (1959), Richter (1966), and Sen (1971). Samuelson sought to reformulate the positive theory of consumer choice so as to eliminate reliance on a subjective notion of preference. His motivation appears to have been philosophical. The empiricism (see §A.1) prevalent in the 1930s made reference to subjective preferences methodologically suspect. Apart from some technicalities, Samuelson succeeded in showing that if choices among commodity bundles satisfy a consistency condition, then a complete and transitive preference ranking can be constructed from the choices. Preferences can be "revealed" by choices, and the empirical legitimacy of talk of preferences can be secured by reducing it to talk of observable choices. In this work, Samuelson is concerned with the positive theory of choice, not with the normative theory of rationality, but his results apply to both. For further discussion of Samuelson's methodological views, see Section 11.2.

The basic idea of revealed preference theory is that, if Mimi chooses option x, when she might have chosen option y, then she has revealed that she prefers x to y or is indifferent between them. Her choices are consistent if they satisfy the "weak axiom of revealed preference" (WARP). It says that if x and y are both in the set of alternatives among which Mimi chooses, and she chooses x, then she never chooses only y from any set including both x and y. In consumer choice theory, the statement of WARP is somewhat more complicated, because prices influence choices by determining which bundles of commodities are available rather than by influencing preferences.

If choices satisfy sufficiently strong consistency conditions, then, in principle, economists can construct a complete and transitive revealed preference ordering from them (Sen 1971, 1973). Samuelson's hope was to purge economics of unobservable and hence (in his view) unscientific content by replacing the axioms governing subjective preferences with an axiom requiring consistency of choice.[5] His view is still popular. For example, in an influential essay, Faruk Gul and Wolfgang Pesandorfer write, "[i]n the standard approach, the terms 'utility maximization' and 'choice' are synonymous" (2008, p. 7).

In fact, revealed preference theory mischaracterizes the notion of preferences that economists employ. Economists do not and cannot employ a notion of preference defined in terms of choices. Economists in fact employ a conception of preferences as subjective states that determine choices only in conjunction with beliefs.

This argument may appear beside the point to economists, who often take "revealed preference" to mean nothing more than inferring preferences from market data given often implicit assumptions about people's beliefs. For example, Boardman et al. write that "[t]he indirect market methods discussed in this chapter are based on *observed behavior*, that is, *revealed preference*" (2010, p. 341). No one doubts that claims about preferences are inferred from behavior (including verbal behavior) and assumptions about beliefs. If only that were all that is meant by speaking of revealed preference theory. Samuelson is after bigger game.

The central claim of revealed preference theory can be formulated as: *A* prefers *x* to *y* if and only if *A* sometimes chooses *x* from

[5] If choice reveals preference, then it is impossible for preferences to be incomplete. Since people always do *something*, even if it is refusing to make a choice, they always reveal a preference. This implication of revealed preference theory violates ordinary usage, but defenders of revealed preference theory need not conform to pretheoretical talk. There are, however, costs. Taking choices to be revealed preferences leads to intransitivities under conditions of risk and uncertainty that have nothing to do with irrationality. If the differences between choices in a sequence $x_1, \ldots, x_n$ are not noticeable, for all i, individuals may be indifferent between x_i and x_{i+1}, but not indifferent between x_1 and x_n.

sets of alternatives that include y, and A never chooses y from any set that includes x. Many economists mistakenly believe that this claim has been proven. For example, Henderson and Quandt write, "the existence and nature of her [an agent's] utility function can be deduced from her observed choices among commodity bundles" (1980, p. 45).

The theorem that Henderson and Quandt have in mind is the following. Suppose that R is a two-place relation such that for some set of alternatives S, available to an individual, Jeff, xRy if and only if Jeff chooses x from S that includes y – that is, if and only if x is in $C(S)$, the set of choices that Jeff makes when he repeatedly chooses from S.

> The *revelation theorem*: WARP implies that R is complete and transitive and the set of maximal elements of S according to R, $\mathbf{Max}^R(S)$, is identical with $C(S)$.[6]

"R" is supposed to be interpreted as "weak preference" ($x \succeq y$). If Jeff weakly prefers x to y, then he satisfies the WARP if and only if Jeff's choice set for any set of alternatives including both x and y never includes y unless it also includes x. The revelation theorem

[6] Here is a sketch of the proof. Let S be a nonempty set of alternatives available to an agent A and $C(S)$ the nonempty subset of S consisting of all the alternatives in S that A actually chooses. Define such at xRy if and only if there is some set S containing x and y for which x is in $C(S)$. The task is to prove that R is (1) complete, (2) transitive, and (3) for any set S, x is in $C(S)$ if and only if, for all y in S, xRy.

(1) Because $C(S)$ is not empty, for all x, y, either x is in $C(\{x,y\})$ or y is in $C(\{x,y\})$ or both x and y are in $C(\{x,y\})$. So either xRy or yRx or both, and R is complete.

(2) Suppose that xRy and yRx. Given the definition of R and WARP, xRy implies that there is no set of alternatives whose choice set includes y, but not x, and yRz implies that there is no set of alternatives whose choice set includes z but not y. So xRy and yRz jointly imply that $C(\{x,y,z\})$ (which is by definition nonempty) consists either of $\{x\}$, $\{x,y\}$, or $\{x,y,z\}$, and all three of these possibilities imply xRz. So R is transitive.

(3) If $x \in C(S)$, then by the definition of R, for all $y \in S$, xRy. Conversely, if for any S, x is not in $C(S)$, then since the choice set is nonempty, for some z it is not the case that xRz. So x is in $C(S)$ if and only if it is in the set of those alternatives in S that are maximal with respect to R. In other words, x is in $C(S)$ if and only if for all y in S xRy.

establishes that if Jeff's choices satisfy WARP, then there is a relation R that is complete and transitive and that implies Jeff's choices. In other words, Jeff acts as if maximizing R.

On the intended interpretations, the revelation theorem establishes that preferences can be defined in terms of choices when choice behavior satisfies WARP. Some economists take the revelation theorem to show that economists can dispense with the notion of preference. On this view, the theorem shows that anything economists need to say about the behavior of individuals can be said in the language of choice (Mas-Colell et al. 1995, p. 5). Other economists regard the correspondence between choice and preference as legitimating talk of subjective preferences. In Sen's words, "[t]he rationale of the revealed preference approach lies in the assumption of revelation and not in doing away with the notion of underlying preferences" (1973, p. 244).

These interpretations of the theorem are not defensible. The binary relation that the revelation theorem proves to be implicit in choices that satisfy the WARP is not the preference relation and cannot serve the functions that the preference relation serves in economic theory and practice. The identity between $\mathbf{Max}^{R}(S)$ and $C(S)$ does not reveal "underlying preferences." Talk of preferences cannot be eliminated from economics without gutting the discipline.

Among the many objections to revealed preference theory,[7] two stand out. First, if preference is defined by choice, then where there is no choice, there is no preference. Revealed preference theory limits preferences to those alternatives among which agents choose. It thus denies that an agent has preferences among infeasible alternatives or among of states of affairs among which the agent faces no choice. Restricting preferences to those alternatives among which people have chosen would cripple economics.[8] Nothing could be said about

[7] For other criticisms see Sen 1971 and 1973. For example, if people choose only a few times from $\{x,y\}$, how can one distinguish preference from indifference? How can one distinguish indifference from violations of WARP or changes in taste?

[8] For example, revealed preference theory implies that indifference curves, which are discussed in Chapter 2, do not exist.

how preferences among the consequences of choices affect choices, because preferences are limited to the objects of choice themselves. The only thing economists could say to predict an agent's choice would be that the agent chooses whatever the agent has chosen.

The obvious response to this serious problem is to reinterpret the theory. Rather than maintaining that an agent such as Jessica prefers x to y if and only if she never chooses y when x is available, revealed preference theorists might say that Jessica prefers x to y if and only if she *would* never choose y if x *were* available (Binmore 1994). On this interpretation of revealed preference theory, whether agents actually face a choice between x and y is irrelevant to their preferences, which are defined by how they *would* choose, if they were to face such a choice.

In switching from actual to hypothetical choice, economists abandon the empiricist project of avoiding references to anything that is not observable. How King Charles would choose if it were up to him whether the USA remains in NATO is no easier to observe than his preference. Hypothetical choices are not choices. They can be predicted, but not observed. Predictions about how Charles would choose rely on no different or better evidence than claims about what he prefers. Notice, in addition, that claims about what he would do in a hypothetical situation cannot be answered until his beliefs are specified. Suppose that Charles were given an apparatus with a blue button that keeps the USA in NATO and a red button that leads it to leave. Without knowing what Charles believes about the buttons, we cannot predict what he would do.

The second problem with revealed preference theory, whether it attempts to define preference in terms of actual or hypothetical choices, is that its fundamental claim is false. It is not the case that if Martha prefers x to y, then she never chooses y or would never choose y, when she could have chosen x. If Martha mistakenly believes that x is not among the available objects of choice, then she may choose y despite preferring x to y. For example, at the end of *Romeo and Juliet*, Romeo enters the tomb of the Capulets and finds

Juliet apparently dead. He does not know that she took a potion that simulates death. Unwilling to go on living without Juliet, Romeo takes poison and dies. He chooses death from a set of alternatives that in fact includes eloping with Juliet. If choice defines preference, then Romeo prefers death to eloping with Juliet. In fact, of course, he prefers eloping with Juliet to death and chooses death only because he does not know that eloping with Juliet is a (so-to-speak) live option.

Defenders of revealed preference might respond as follows:

> The second criticism shows only that beliefs mediate between choices and preferences, when preferences are understood as they are in everyday conversation. In contrast, in economics, as the revelation theorem shows, consistent choice demonstrates the existence of a complete and transitive relation that gives a top ranking to the alternatives individuals choose. This relation, call it "preference*," is the preference relation that economists rely on, and it is provably derivable from choice. Unless Romeo violates WARP – and given the nature of his choice, his future consistency is guaranteed – his choice reveals his preference* for death over eloping with Juliet.

On this view, economists employ a technical concept, preference*, that is defined in terms of choice. It is unfortunate that their use of the same term confuses outsiders, but the economist's notion of preference* is defined by choice.

Economists are entitled to define their own technical concepts and to proscribe the use of everyday concepts, but only if they actually use the concepts they define rather than the concepts they proscribe. In fact, economists rely on a concept of preferences that is not revealed by choices and they cannot avoid doing so without eviscerating their theories. For example, when Donald Trump was elected, the price of stock in private prisons, which Hillary Clinton had proposed shutting down, shot upward. To explain and predict this, economists need to cite the beliefs of investors as well as their preference for higher returns. But earning

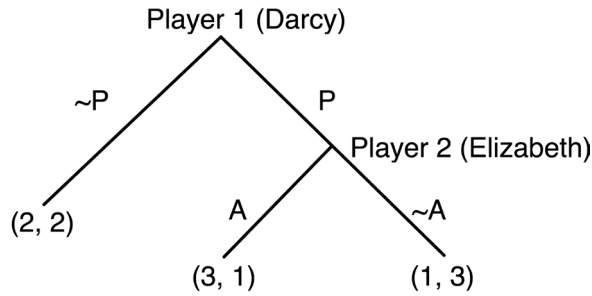

FIGURE 1.3 Darcy and Elizabeth.

a higher return is not an object of choice, and the preference for higher returns is not a revealed preference. Preferences, as understood by economists, explain behavior only in conjunction with beliefs.

Moreover, if economists took preferences to be revealed preferences, they could not do game theory. Consider, for example, the scene from *Pride and Prejudice* where Darcy, overcome by his love for Elizabeth, proposes marriage to her, despite her lack of dowry, her mother's vulgarity, and her younger sister's silliness and impropriety. Regarding Darcy as arrogant and unfeeling, Elizabeth turns him down. Their interaction can be modeled as a game (Figure 1.3).

The numbers in Figure 1.3 are ordinal utilities – that is, indicators of preference order. Higher numbers indicate more preferred alternatives. The first number in each pair expresses Darcy's utility, and the second number expresses Elizabeth's utility. Darcy moves first and can either propose (P) or not propose (~P). Not proposing ends the game with the second-best outcome for both players.[9] If Darcy proposes, then Elizabeth gets to choose whether to accept (A) or reject his proposal (~A). Rejecting the proposal is the best outcome for Elizabeth (at this point in the novel) and the worst for Darcy, while accepting is best for Darcy and worst for Elizabeth.

[9] It is arguable whether Elizabeth preferred to receive and reject Darcy's proposal over not receiving the proposal. Whether I am right about Elizabeth's preference does not matter to the point the example makes.

Some of the preferences in Figure 1.3 are revealed by choices. For example, Elizabeth's refusal reveals that she prefers rejecting to accepting the proposal. However, other preferences, which are needed to define the game, rank alternatives between which agents do not and cannot choose. For example, Darcy cannot choose whether Elizabeth accepts, but the game is not well defined without specifying his preference over her acceptance or rejection. To predict whether Darcy will propose, a game theorist needs to know Darcy's preferences among the outcomes, including outcomes between which he cannot choose, as well as his beliefs about whether Elizabeth will accept his proposal. Preferences in games are not preferences* (Rubinstein and Salant 2008, p. 119).

Beliefs mediate the relationship between choices and preferences. Economists can infer preferences from choices or choices from preferences only given premises concerning beliefs. Neither beliefs nor preferences can be identified from choice data without assumptions about the other. Choices can be evidence of preferences, but they cannot define them.[10]

Economists have paid little attention to these objections because they often restrict their models to circumstances where what people believe coincides with what is truly the case. If beliefs match the reality, then economists need not mention them. That fact makes beliefs no less important.[11] Preferences cannot be defined by choices, because preferences cannot be limited to the immediate objects of choice and because they cannot be inferred from choices without premises concerning beliefs.

[10] For just one example of what this means in practice, consider the study carried out by Henderson et al. (2011). On the basis of data concerning how much Kenyans are willing to pay to protect their water sources, Henderson et al. calculate their willingness to pay for protecting their children from diarrhea. These inferences depend on what the Kenyan parents *believe* about the effects of protecting water sources and the causes of diarrhea.

[11] To defend revealed preference theory, Johanna Thoma (2021a) proposes that economists can take beliefs to define the objects among which individuals choose, which would then permit them to identify preferences with choices. But in that case, choices themselves are not observable without information about the agent's beliefs, and the proposal does nothing to mitigate the other objections to revealed preference theory.

1.3 RATIONALITY AND UNCERTAINTY: EXPECTED UTILITY THEORY

The theory of rationality can be extended to choices involving risk and uncertainty. Economists and decision theorists commonly speak of *risk* when agents know the possible outcomes of their choices and their probabilities. In situations involving *uncertainty*, it is not known what are the probabilities of the outcomes of the alternatives or even what the outcomes may be.[12] I treat the cases of risk and uncertainty together by allowing the probabilities mentioned in Section 1.3.1 to be either limits to relative frequencies or subjective degrees of belief. This simplification begs the question against those who maintain that situations of uncertainty involve more radical ignorance and different principles of rational decision-making.

1.3.1 Conditions on Choice When There Is Risk or Uncertainty

An action whose outcome is not known can be treated as if it is a *lottery* with its possible outcomes as the prizes. For example, suppose that Amy has the option of approaching a lost dog in the hope of returning it to its owner. She does not know what the outcome will be, but she thinks there are three possibilities: it runs away with or without biting her first, or she succeeds in returning it. The subjective probability or degree of belief that Amy attaches to the three outcomes are: Pr(dog runs away without biting her) = 0.3; Pr(dog runs away and bites her) = 0.1; and Pr(Amy returns dog to owner) = 0.6. The alternative of approaching the stray can then be represented as a lottery with three prizes that occur with the respective probabilities. Explaining or predicting what Amy winds up doing requires knowing not only her subjective probabilities but also her preferences among the alternatives. If she cares much more about whether she is bitten

[12] See Luce and Raiffa 1957, chapter 2. Some Bayesians (§A.7) deny that there are such things as objective probabilities. Recently decision theorists have used "ambiguity" to refer to what I called "uncertainty."

than whether she gets the dog back to its owner, then despite the low probability of getting bitten, she will not approach the stray.

One can represent lotteries as a pair $[R, p]$, where R is a set of mutually exclusive and jointly exhaustive pay-offs, and p a probability measure defined on R. The lottery that pays off K with probability p and L with probability $(1 - p)$ can be denoted conveniently as $[K, L, p]$ or as $\left[(K, p), (L, 1 - p)\right]$. Since the choice of an action that leads with certainty to a particular outcome K can be represented as a "degenerate" lottery $\left[(K, p), (K, 1 - p)\right]$ or as $\left[(K,1), (x,0)\right]$, one can without loss of generality conceive of all the objects of preferences as lotteries. These lotteries include alternatives such as bets on ball games, where the probabilities are subjective degrees of belief. One should not be misled by the lottery terminology. Economists set aside (via "the reduction postulate") the pleasures of gambling.

In offering a normative theory of decision-making under risk and uncertainty, economists assert – as before – that preferences (whose objects are now conceived of as lotteries) are complete, transitive, reflexive, continuous, and stable. In addition, one needs a "reduction postulate" relating compound and simple lotteries. Harsanyi calls it a "notational convention" (1977b, p. 24), and it serves as a criterion of identity for lotteries. For example, suppose Peter faces the following compound gamble: if a coin comes up heads, then he can roll a die and win \$7 if the die comes up 6 and \$1 otherwise. If the coin comes up tails, he draws from an urn containing three red balls and one white ball, winning \$7 if he draws a red ball and losing \$1 if he draws a white ball. The reduction postulate says that this complex lottery, $\left[(\left[(\$7, 1/6), (\$1, 5/6)\right], ½), (\left[(\$7, ¾), (–\$1, ¼)\right]), ½)\right]$, is equivalent to the simple lottery one gets when one substitutes for the embedded lotteries their expected values – in this case $\left[(\$2, ½), (\$5, ½)\right]$, which looks like it would be less fun than the gamble Peter faces. The reduction postulate implicitly rules out preferences for gambling itself.

Expected utility theory, the theory of rationality under circumstances of risk and uncertainty, relies on one other substantial and important axiom, called the "independence" condition or "the

sure-thing" principle. It should not be confused with the context independence discussed earlier. The independence principle says that, if two lotteries differ only in one prize (which may itself be a lottery), then preferences between the two lotteries should match preferences between the prizes: If $L_1 = [(x, p), (y, 1-p)]$ and $L_2 = [(z, p), (y, 1-p)]$, then the independence axiom states that A prefers L_1 to L_2 if and only if A prefers x to z.

1.3.2 *The Cardinal Representation Theorem*

Given completeness, transitivity, reflexivity, continuity, the reduction postulate, and the independence principle, it is possible to prove a (cardinal) representation theorem, which is much stronger than the ordinal representation theorem discussed in Section 1.1.3:[13]

> If all of these axioms are true of an agent's preferences, then those preferences can be represented by a utility function with the expected utility property, which is unique up to a positive affine transformation.

A utility function possesses the expected utility property if and only if the (expected) utility of any lottery is equal to the utilities of its outcomes weighted by their probabilities, for example $U([(K, p), (L, 1-p)]) = pU(K) + (1-p)U(L)$. A positive affine transformation of an expected utility function U is a linear function $aU + b$, where a is a positive real number and b is any real number. The representation theorem establishes that if an agent's preferences satisfy all the conditions, then the agent's expected utilities are as measurable as temperature is on the centigrade or Fahrenheit scales. The zero point and units in an expected utility scale are arbitrary, but nothing else about the scale is. Comparisons of utility *differences* are independent of the scale chosen. If $U(x) - U(y) > U(z) - U(w)$,

[13] For an accessible presentation, see Harsanyi 1977b, chapter 3. Other proofs can be found in Herstein and Milnor 1953, Jensen 1967, and von Neumann and Morgenstern 1947.

and U' is a positive affine transformation of U, then $U'(x) - U'(y) > U'(z) - U'(w)$.[14]

As in ordinal utility theory, economists assume choice determination: among the alternatives that agents believe to be feasible, agents choose an alternative at the top of the ranking. When economists speak of agents "maximizing utility," this is what they mean – nothing more. Utility is still only an indicator of preferences, although now it indicates preference intensity as well as preference order.

If the axioms of expected utility theory are true of an agent A and A's preferences are stable, it is in principle possible to determine both A's utility function and A's probability judgments by observing A's choices among lotteries. For example, suppose that, as in Figures 1.1 and 1.2, Marianne prefers bread to bananas and pineapple, among which she is indifferent, and that she prefers bananas and pineapple to carrots and carrots to apples. Since the zero point and the units of her utility function are arbitrary, one can stipulate the values for utility of an apple $U(A)$ and the utility of bread $U(B)$. Given these axioms, for some probability p, Marianne will be indifferent between pineapple for certain and a lottery that pays off bread with probability p and an apple with probability $1-p$ (that is, the lottery [(bread, p), (apple, $1-p$)]). The utility of a pineapple, $U(P)$ will then equal $pU(B) + (1-p)U(A)$. The probability an agent attaches to an event E can be determined when one knows the expected utilities of a lottery and its prizes when the prizes depend on whether E occurs.[15]

The probabilities invoked in such an elicitation process are personal subjective probabilities, that is, the degrees of belief of individuals; and the axioms for rational choice under conditions of

[14] This is easily proven. Suppose (1) $U(x) - U(y) > U(z) - U(w)$, and (2) $U(.) = aU'(.) + b$, where $a > 0$. Substituting $aU'(.) + b$ for $U(.)$ gives us (3) $aU'(x) + b - aU'(y) - b > aU'(z) + b - aU'(w) - b$. The b's cancel out, and since a is positive, one can divide through without changing the sign of the inequality. Thus (4) $U'(x) - U'(y) > U'(z) - U'(w)$.

[15] Given my short-cut description, it might appear that one cannot elicit *both* probability judgments and a utility function. But (although not without a further assumption) one can – see Ramsey 1926.

uncertainty imply that these degrees of belief must satisfy the axioms of the probability calculus. Moreover, if Greg's degrees of belief do not satisfy the axioms of the probability calculus, then Greg can be led to accept a series of bets on some chance event *E*, leading to a certain loss whether *E* occurs or not. This demonstration is known as the "Dutch Book argument" (see Schick 1986 for a critical discussion). Expected utility theory is a theory of rational belief as well as a theory of rational preference and choice. Subjective probabilities may arise from knowledge of objective frequencies, but they need not. The formal theory of choice is itself silent on the origin and justification of probability judgments. Those who have made the most of this theory, so-called personalist Bayesian philosophers and statisticians, are permissive about the grounds for these probability judgments.

1.3.3 Expected Utility Theory and Its Anomalies

In summary, expected utility theory, as a theory of rationality, can be presented as follows:

1. An agent *A*'s *choices* are *rational* if and only if: (a) *A*'s preferences and beliefs are rational and (b) *A* prefers no option to the one *A* chooses among the options that *A* believes to be feasible.
2. An agent *A*'s *preferences* are *rational* if and only if:
   a. *A*'s preferences are complete, transitive, reflexive, and continuous,
   b. *A* is indifferent between options the reduction postulate identifies, and
   c. *A*'s preferences satisfy the independence condition.
3. An agent *A*'s degrees of belief are rational if and only if they satisfy the axioms of the probability calculus.

Expected utility theory is a stunning intellectual achievement, which forms the foundation for contemporary decision theory. Although it often puts in an appearance in economics, it is not nearly as important to day-to-day economic theorizing as ordinal utility theory.

Unlike ordinal utility theory, which is testable only in the unusual circumstances in which there is perfect knowledge and no uncertainty, expected utility theory purports to apply to ordinary

decision contexts both as a source of predictions concerning what people will choose (if they choose rationally) and as a source of normative recommendations concerning what choices are rational. Economists and psychologists can study whether people actually choose the option that expected utility theory says they do and should. Claims about how people actually choose are much more easily testable than claims about how they should choose. Investigations showing that the predictions of expected utility theory are not borne out might only show that people fail to choose rationally. But it is important to assess the normative adequacy of both ordinal utility theory and especially expected utility theory, because they claim to guide decision-making. They matter. The account of rationality one relies on influences policy-making. Although the issues are highly theoretical, their resolution is deeply practical.

What are the issues? First, questions concerning completeness, independence, and continuity become more troubling once uncertainty is admitted. When individuals are unable to rank options, is the uniquely rational response to make guesses about the probabilities of outcomes in order to compute expected utilities? Why should a rational agent's ranking of two lotteries K and L never be affected by the discovery of other options? Continuity implies that, if a rational individual Arlo prefers \$100 to \$10 and \$10 to slow fatal torture, then there is some probability p less than one such that the lottery that pays off \$100 with probability p and slow fatal torture with probability $1 - p$ would be worth at least \$10 to Arlo. Is he irrational to refuse to accept this lottery?

The new axioms that expected utility theory adds to ordinal utility theory are problematic, too. The reduction postulate is questionable, because there seems to be nothing irrational about someone who enjoys gambling preferring a compound lottery to the simple lottery to which it reduces.[16] Although controversy concerning

[16] Perhaps one might regard the reduction postulate, like completeness, as narrowing the domain to which expected utility theory applies.

expected utility theory has focused on the independence condition, it actually seems at first glance easier to defend. In the case of indifference, it serves as a substitution principle. If agents are indifferent between options x and y, then substituting one for the other in a gamble should make no difference. When there is a strict preference, the independence principle seems to follow from considerations of dominance. Suppose, for example, that lotteries K and L involve flipping a coin. If the coin comes up heads, K has a better prize than L, while the prizes if they come up tails are the same. One can do no worse with K and may do better. On the basis of an argument like this one, Savage called a version of the independence principle the "sure-thing" principle (for a simple exposition see Friedman and Savage 1952, pp. 468–9).[17]

Yet, many have found the independence condition unacceptable. As the case study in Chapter 14 illustrates, there are instances in which individuals not only seem to violate it, but in which the violations appear to be rational. Echoes of the controversies concerning expected utility theory are heard within economics, but less often than one might expect, because economic models so often employ only ordinal utility theory. The challenges to expected utility theory raise interesting methodological issues about the role of evidence in economics, which I discuss in Chapters 15 and 16, but I do not attempt to resolve the deep problems concerning the nature of rationality touched on earlier.

1.4 WHAT ARE PREFERENCES?

The discussion of the axioms of ordinal and expected utility theory, the implicit assumptions concerning preferences, and the mistakes of revealed preference theory jointly pin down the conception of

[17] This reasoning supposes that the choice of L rather than L^* does not affect the value of P, and it does not necessarily carry over to the case where the prizes in the lotteries are themselves lotteries.

preferences that lies at the heart of mainstream economics.[18] One can read off an interpretation of preferences from the following assumptions about preferences: preferences are (at least to some degree of approximation) complete, transitive, reflexive, and continuous; and they satisfy the independence condition. They are given and largely stable over time and across contexts, and the alternatives that they rank are complete states of the world. These assumptions imply:

1. *Preferences are comparative evaluations.* They are evaluative, because they can be expressed in the form of a ranking in terms of better or worse. They are comparative: to say that Mary prefers to go dancing is elliptical. She prefers dancing to something else.
2. *Preferences are "total" comparative evaluations that motivate choices*. They rank states of affairs, including the immediate objects of choice, as better or worse with respect to everything the agent considers to be relevant. Note that I make no assumption concerning *what* the agent considers to be relevant, nor concerning whether the agent is rational or well informed concerning her judgment of what is relevant to a choice. An agent's preference ranking may depend on a few largely irrelevant properties of alternatives, or it may reflect an exhaustive investigation of the options.
3. *Preferences are subjective states that determine choices* in combination with beliefs and constraints. As subjective states, they are not directly observable. They can be inferred from choices – but only with the help of premises concerning beliefs.
4. *Preferences are subject to rational criticism.* They are not just gut feelings, even if sometimes they depend on nothing else.

Preferences must be total evaluations (point 2) because in combination with beliefs and constraints, they determine choices. They thus cannot be "partial" comparative evaluations of alternatives. From the agent's perspective, preferences rest on a comparison in every

[18] I have in mind the preferences of human economic agents. It is also possible to talk about the preferences of groups, animals, plants, and even machines; and one may want to make different claims about preferences of other sorts of agents. See Guala 2019.

relevant regard. I take it as implicit in the notion of an evaluation that it motivates choices. As total comparative evaluations, preferences in economics differ from preferences in everyday conversation, in which obligations and commitments *compete* with preferences in determining choices and the value of alternatives. Whereas noneconomists might say, "Bonnie preferred to go out with her friends to staying home; nevertheless, she stayed home because she promised to babysit," economists would say that Bonnie preferred to stay home because she promised to babysit. In economic models of rational choice, whatever influences choices, other than beliefs and constraints, does so via influencing preferences.

More should also be said about the vulnerability of preferences to rational criticism, because many economists have denied it. Although in their famous paper "De Gustibus Non Est Disputandum" ("There is no arguing about tastes") (1977), George Stigler and Gary Becker deny that preferences among commodity bundles should be regarded as primitives in economics, beyond explanation, they attribute to most economists the belief that "when a dispute has been resolved into a difference of tastes," "there is no further room for rational persuasion" (1977, p. 76). They are right that such a view is prevalent among economists. Nevertheless, it is mistaken. Although Margaret may regard taste as the only factor that is relevant to her preference for a strawberry ice cream cone over a coffee ice cream cone, even a preference such as this one lays hostages to rational criticism. A newspaper article concerning an *E. coli* outbreak caused by eating strawberry ice cream may change Margaret's preferences.[19] With new experiences and information, she may change the list of factors that she considers to be relevant to her preference. Satisfying the axioms of ordinal or cardinal utility theory can sometimes be a demanding cognitive task. Mas-Colell et al. maintain that "[i]t takes

[19] One might instead maintain that the newspaper article leads Margaret to believe that she was mistaken about which alternatives her (unchanged) preferences rank. Her choice of ice-cream flavors is nevertheless subject to rational criticism whether one takes the new information as changing preferences or changing the alternatives.

work and serious reflection to find out one's own preferences" (1995, p. 6). In short:

> *Preferences are total subjective comparative evaluations*, which are *subject to rational criticism*.

The models of rational and actual choice employed by economists explain and predict behavior by citing the constraints on choices and the agent's beliefs and preferences. Constraints on choices typically limit choices via beliefs. People who are late to an appointment do not flap their arms in a futile attempt to fly. Because they know that flying unassisted is not possible, they do not try. The axioms concerning preferences say nothing about *what* people prefer. Unusual people, who long for pain and suffering, could satisfy the axioms. Positive economic theory supplements the axioms of ordinal utility theory with axioms concerning the content of preferences, such as the claim that people prefer more commodities to fewer. These additional axioms are among the subject matter of Chapters 2 and 3.

1.5 PREFERENCES AND SELF-INTEREST

Neither ordinal utility nor expected utility say anything about the extent to which individuals are self-interested. However, the fact that the standard models of rational choice take an agent's choices to be determined by the agent's own preferences has misled economists and commentators on economics into thinking otherwise. Even the Nobel laureate, Amartya Sen, has on occasion mistakenly taken preference to imply self-interest. He maintains that "preference in the usual sense" has "the property that if a person prefers x to y then he must regard himself to be better off with x than with y" (1973, p. 67). "Preference can be … defined so as to keep it in line with welfare as seen by the person in question" (1973, p. 73), and "the normal use of the word permits the identification of preference with the concept of being better off" (1977, p. 329). Similarly, Daniel Kahneman maintains that economists typically equate what people choose

with what they anticipate will result in the most enjoyment (2006, pp. 489, 501).

Self-interest or expected advantage cannot be what people *mean* by preference, because there is no contradiction in maintaining that people's preferences may depend on things that people do not expect to influence their own well-being. Most people do not apportion their donations to disaster relief by considering how much those donations will contribute to their own well-being. Drivers in the grip of road rage, who have shot and killed other drivers, are focused on harming others rather than benefiting themselves. Consider the humdrum instrumental decisions that fill one's life. People often have no idea how they bear on their interests. When deciding among shoes for a seven-year-old, parents are thinking about which pair would be best for the seven-year-old, not for themselves. The mere *possibility* that people have preferences among alternatives, without considering how they influence their own interests or that people sometimes sacrifice their interests in order to accomplish something that matters more to them, shows that doing as one prefers is not by definition acting in one's self-interest or promoting one's expected benefits.

And these are not mere possibilities: apart from sociopaths, people are capable of distinguishing what they want most of all from what they judge to be best for themselves, and most people sometimes carry out actions whose consequences they believe to be worse for themselves than some feasible alternative. Moreover, if, as many welfare economists assume, well-being is defined as preference satisfaction, then preferences cannot be defined by expected well-being.

What leads to the conflation of preference and self-interest is that one's preferences reflect one's interests, and speaking of acting on one's interests invites an equivocation between acting "in pursuit of one's objectives (whether self-benefiting or not)" and acting "in pursuit of one's own advantage." There may be some individuals whose objectives are limited to benefiting themselves. But most people have all sorts of objectives. The pursuit of some project that is not intended to benefit oneself may of course wind up benefiting oneself.

Indeed, venerable advice for living well counsels devoting oneself to something other than one's own interests. But there is nothing in this good advice that equates preference and self-interest.

Many economic models take people to be self-interested, and for specific purposes, such models are often useful. I would be skeptical of a model of private equity companies that attributes to the executives of those firms entirely altruistic preferences. But self-interest is not built into the meaning of preferences. Utility theory places no constraints on what individuals may want; it only requires consistency of preferences and that choices manifest preference, given belief. Utility theory has a much wider scope than economics. As is appropriate in a theory of rationality, it says nothing specifically about commodities or services. It says nothing about people's aims, about whether agents are acquisitive and self-interested or generous and otherworldly, or about whether humans are saints or sinners.

1.6 CONCLUSIONS

Mainstream economists employ a model of rational choice, which they also take to be an approximate characterization of actual choice. In this model, choice is determined by constraints, beliefs, and preferences. While not providing an explicit definition of preferences, economists are committed to a set of axioms and standard assumptions concerning preferences that together imply that preferences are total subjective comparative evaluations. Preferences are not beyond criticism, nor is it the case, as some economists have maintained, that economists have nothing to say about their formation and modification. Ordinal utility theory is a convenient way of expressing the consequences of the conditions economists impose on choices and preferences (and, in the case of expected utility theory, beliefs as well). As Chapters 2 and 3 show, this model of rational choice is embedded in microeconomics, general equilibrium theory, and macroeconomic models.