Chapter 17

A Difficult Choice in Preference Theory: Rationality Implies Completeness or Transitivity but Not Both

Michael Mandler

1 INTRODUCTION: RATIONALITY IN PREFERENCE THEORY

The economic theory of rational choice enjoys ever-widening popularity. Various social sciences now routinely endow agents with preference orderings or utility functions and explain social outcomes as the product of maximizing behavior. Curiously, the boom in rational-choice theory outside economics has coincided with growing doubts about the theory on the home turf. After long ignoring the substantial evidence that individuals do not choose as theories of rationality wish them to, economists have increasingly turned to positive models of choice, often derived from psychology, that make no mention of ideal or rational conduct (see Rabin 1998 for a recent survey).

At first glance, it seems odd that economics ever aspired to a normative theory of rationality. A science in the business of prediction can seemingly ignore the question of how agents ought to choose, and thus sidestep the controversies that inevitably surround definitions of what is rational. The risks of embracing a dubious theory of rationality are not mere abstract possibilities. For decades, economics has been taken to task for claiming that agents are self-interested pleasure seekers. Partly as a reaction, the economic theory of rationality has evolved considerably over the last hundred years. Originally, to be rational was indeed to choose options that deliver the greatest pleasure. But at least since the 1930s, rationality in economics has been identified instead with the more modest standard that preference be internally consistent; agents in economics no longer pursue the fictional substance called utility. This shift, which remains underappreciated outside of economic theory, is one key to why rationality has remained central to preference analysis. When

narrowed to internal consistency, rationality seems to place only weak plausible restrictions on behavior.

The claims of preference theory are also less ambitious than is sometimes supposed. Economic analysis does not assert the absurdity that agents always choose the preference-maximizing action. The theory claims only that when agents systematically violate the dictates of economic rationality—which posit that agents can rank any pair of options and that rankings are transitively ordered—they suffer harm. Consequently, given practice and opportunity to learn, their behavior will in time conform more closely to the axioms of rationality. For many, this long-run link to behavior explains the role of rationality in preference theory: rationality can ultimately guide action.

But despite the common belief that the axioms of economic rationality are incontestable features of reasonable conduct, preference theory does not adequately explain why behavior should obey those axioms. Instead, the axioms of rationality have taken on a life of their own. In the absence of clear justifications for the rationality axioms, the behavioral evidence that contradicts these axioms is difficult to assess. Does the problem lie with the behavior or the axioms? That is, are agents indeed acting self-destructively, or do the axioms of rationality mischaracterize which patterns of choice are reasonable?

My primary aim in this essay is to show that it is the axioms that are to blame. To accomplish this, I reconstruct those partial arguments in favor of the rationality axioms that do exist. As we will see, there are strong cases for the rationality of the completeness axiom—the assumption that agents can rank any pair of options—and for the rationality of transitivity, but the arguments in favor of each axiom employ different definitions of preference. Completeness applies to preference as choice, while transitivity applies to preference as a set of judgments of well-being. Convincing arguments for the axioms taken together cannot be assembled on either definition.[1]

I distinguish between preference theories that put forward *ordering principles*, which explain how agents come to their preference rankings, and those that do not. The hedonism advocated by the inventors of economic utility theory was decidedly in the first camp. Such theories give grounds for why agents should have well-defined judgments about what promotes their welfare; that is, they explain why preferences in the welfare sense should be complete. In view of hedonism's manifest implausibility as a theory of motivation, its expulsion from economics has seemed an

unqualified gain. But the need for ordering principles remains, although nowadays it is rarely acknowledged. As we will see, the difficulties of current-day preference theory stem from its attempt to impose completeness and transitivity as universal axioms, when in fact their plausibility hinges on whether or not an ordering principle is present.

To illustrate the role of ordering principles in preference analysis, I begin the essay with a brief look at hedonistic preference theory. Conveniently, this will allow a detour to the theory of cardinal utility, which is the natural model of utility for pleasure-seeking agents. I then turn to the movement that overturned hedonism and cardinal utility, ordinal preference theory, which remains the cornerstone of preference analysis to this day. After exploring my central topic—the limits of ordinalism's ability to defend its account of rationality—I return to cardinality. As we will see, cardinal utility provides an ordering principle for the theory of choice under uncertainty. Analogously to the difficulties facing standard choice theory in the absence of the ordering principle once supplied by hedonism, the theory of choice under uncertainty cannot easily justify completeness in the absence of cardinality.

Sections 3 and 7 below, on cardinality, are the more technical parts of this essay. They are self-contained and the remainder, with the exception of a stray remark on cardinality in section 4, can be read without them.

2 PREFERENCE BASED ON UTILITY

The theory of utility maximization originally relied on a narrow view of motivation. Particularly in the work of Jevons (1871), one of the founders of neoclassical economics, the only pertinent feature of a good or commodity is the quantity of utility or pleasure it delivers to its consumer. An agent's total satisfaction is the sum of these pleasures across all goods, and agents strive to maximize this sum.

Jevons took utility or pleasure as his primitive concept: utility objectively determines which choices best promote an agent's well-being. Faced with various arrays of goods, it is objectively in the agent's interest to choose the array that delivers the greatest quantity of pleasure. Individuals may on occasion err and fail to choose the array of goods that delivers the greatest pleasure, but with time and leeway for experimentation, individuals will gravitate to the correct, utility-maximizing decision. Utility thus originally served as an ordering principle; it prescribed which choices are rational.

Jevons and other early utility theorists were careful to limit the domain of their analysis to standard consumption goods. They reasoned that material satisfactions, whatever their source, are always commensurable. Shelter from the cold, for instance, delivers the same sort of pleasure-stuff, though a different quantity, as a fine meal. When choices cannot be reduced to homogeneous pleasure—for example, when deciding between altruistic sacrifice and self-interested gain—decision-making cannot proceed via the pleasure calculus, and therefore is not the subject of utility analysis. Jevons and his follower Alfred Marshall took particular care to exempt ethical decisions from the domain of utility theory; hedonism does not supply an adequate ordering principle for such questions. Jevons understood that when decisions do have an ethical dimension, one can still *define* chosen options to embody more pleasure than rejected options. But since it is merely a label, such a concept of pleasure does not prescribe action or determine an ordering; it only certifies after the fact that chosen alternatives have more "pleasure" than rejected alternatives. Jevons consequently rejected this approach.

3 SEPARABILITY AND CARDINALITY

The previous section implicitly treats the pleasure of a good as unaffected by the quantities of other goods consumed. This feature, which I call the *separability postulate*, was an explicit part of the work of Jevons and other early utility theorists. If we let the consumption of good i be denoted x_i and the pleasure or utility of good i as u_i, the separability postulate can be expressed as the assumption that u_i is a function of x_i alone. If there are a total of l goods, the agent's total pleasure or total utility is then $u_1(x_1) + \cdots + u_i(x_i) + \cdots + u_l(x_l)$, which I also represent as $u(x_1, \ldots, x_l)$. Functions u of this mathematical form are called *additively separable*. I will use x as shorthand for a "consumption bundle" of the l goods $(x_1, \ldots, x_l)$.

In current-day preference theory, any *increasing transformation* of $u(x)$, say $F(u(x))$, is considered to be an accurate summary of the agent's preferences. A transformation F is increasing if it satisfies the property: if $u > u'$ then $F(u) > F(u')$. Consequently, if $\hat{x}$ delivers greater utility than x' according to the utility function $u(x)$ and if F is increasing, then $F(u(\hat{x}))$ will be greater than $F(u(x'))$. Evidently, the utility function $F(u(x))$ records the agent's relative ranking of consumption bundles just as accurately as the original utility $u(x)$.

The separability postulate, however, imposes further restrictions on which utility functions constitute fully accurate psychological measuring sticks. Among the functions that can be generated via some increasing transformation F from an additively separable $u(x)$, Jevonian theory effectively deemed only those $F(u(x))$ that preserve the property of additive separability to be acceptable. The other $F(u(x))$, even though they summarize the agent's relative ranking of consumption bundles correctly, fail to record the agent's judgment that the pleasures of distinct goods do not interact with each other.

Consider a couple of examples. Suppose that there are two goods and that $u(x) = x_1 + \log x_2$ is a fully accurate utility function for some agent. That is, $u(x)$ records both the agent's relative rankings of consumption bundles and his or her sensation that goods deliver utility without interaction effects. Since multiplying by 2 is an increasing transformation and preserves additive separability, the utility $2x_1 + 2 \log x_2$ is also fully accurate. But consider instead $(x_1 + \log x_2)^3$. Although cubing is also an increasing transformation, $(x_1 + \log x_2)^3$ does not satisfy additive separability, as the reader can confirm by multiplying this expression out.

These examples hint at a remarkable feature of additively separable utility functions. If we insist that only additively separable utility functions are fully accurate descriptions of some agent's preferences, then we are effectively specifying a cardinal utility function for that agent. To say that utility is cardinal means that if a function u is a psychologically accurate utility function for an agent, then the function v is also psychologically accurate if and only if v is an increasing linear transformation of u. An increasing linear transformation of a function u is a function of the form $au + b$, where $a > 0$. It is easy to see that if u is additively separable, then so is $au(x) + b$, as the case of multiplying by 2 (i.e., $a = 2$, $b = 0$) illustrates. Also, though it is a little trickier to prove this formally, if F is *not* linear, then $F(u(x))$ will be not be additively separable, as the case of cubing illustrates.[2] (Equivalently, if $F(u(x))$ is additively separable, then F is linear.) Thus, if u is additively separable and if we insist that only additively separable functions can serve as accurate utility functions, the entire set of admissible utility functions is precisely the set of increasing linear transformations of u; in other words, utility is cardinal.

Cardinality of utility means that an agent's satisfaction is measurable in the same sense that some physical magnitudes, for example, temperature, are measurable. Specifically, the ratios of utility differences take on a fixed value: given any four consumption bundles x, y, z, w such that z and w do not deliver the same utility level, the ratio

$$\frac{u(x) - u(y)}{u(z) - u(w)}$$

will equal the same number, whichever function u in a set of cardinal utility functions is plugged in. (This fact is easy to confirm: for each appearance of the function u above, simply substitute any given increasing linear transformation $au(x) + b$ and cancel terms.) Cardinality therefore implies that agents can not only judge which changes are more preferred —for example, that a switch from y to x delivers a bigger pleasure boost than a switch from w to z—but can also assign an exact number to the ratio of these changes—the first switch delivers, say, 2.3 times the pleasure of the second. The separability postulate, which at first glance seems to be an innocuous and plausible restriction, ends up implying that pleasure behaves like a tangible, corporeal quantity. Of course, in the nineteenth century this physicality seemed fitting; if homogeneous pleasure is indeed the motivating force behind preference, it is only natural for utility to be cardinally scalable.

4 ORDINAL PREFERENCES AND DERIVED UTILITY

Even restricted to standard consumption goods, hedonism offers a narrow and implausible psychology. While some consumption goods are nothing more than vehicles for pleasure, many are not; they deliver incommensurate benefits and communicate diverse messages. Ways of life require certain commodities; decisions about such commodities cannot be made on the basis of pleasure any more than can the underlying decisions about life. A summer devoted to self-improvement—studying a new language, say—calls for one set of commodities; a summer of fun at the beach calls for another. Yet a decision between scholarship and sunbathing is not made by comparing quantities of homogenous pleasure; it involves judgment of the value of learning, assessment of how to balance recreation and education, awareness of the risks of skin cancer, etc. A fortiori, when we leave the realm of standard goods and consider the intangibles over which preferences are nowadays defined—it is standard, for example, for agents to be endowed with well-defined preferences over the well-being of others—the inapplicability of hedonistic psychology becomes indisputable.

It is therefore unsurprising that economic theory has deserted hedonism wholesale. Faced with criticism of utilitarian psychology, economists began as early as the late nineteenth century to disavow hedonism (Lewin

1996). By the early twentieth century, it had become routine for economists to assert that utility theory was not wedded to any specific psychological model. Utility, economists have claimed ever since, is just a concise way to summarize an agent's relative or ordinal ranking of commodity bundles; it is not supposed to explain how those rankings are psychically crafted.

Ordinal preference theory formalized this new understanding of utility in the 1930s and rapidly achieved theoretical dominance. Current-day economic theories of preference and choice continue to follow ordinalist methodology. The primitive concept of ordinalism is an agent's *preference relation*, usually denoted by the symbol $\succeq$. The expression $x \succeq y$ means that the agent prefers x to y in the weak sense that the agent either strictly prefers x to y or is indifferent between the two. Strict preference and indifference are defined formally in terms of $\succeq$: x is strictly preferred to y, denoted $x \succ y$, if $x \succeq y$ and it is not the case that $y \succeq x$, and x and y are indifferent if both $x \succeq y$ and $y \succeq x$.

Ordinal rankings have two primary interpretations. In the first, to say that an agent strictly prefers bundle x to bundle y, or $x \succ y$, means no more than that the agent systematically chooses x over y. In the second, strict preference for x over y implies in addition that the agent judges him or herself to be better off with x than with y. The first interpretation explicitly avoids psychological content, but even in the second understanding, the meaning of "better off" is intentionally left vague. Economists frequently think of being better off as an experience of greater "well-being" or "welfare," and, for brevity's sake, I will use these expressions too. But the agents of ordinalist theory need not judge what makes them better off by comparing quantities of "welfare." Instead, agents can deliberate about what values take precedence; they may, for instance, reason that religious law rather than sensory pleasure should dictate what foods they eat. That preferred choices deliver greater "welfare" thus means simply that an agent's deliberation has reached resolution. Contemporary preference theory, therefore, is not subject to the criticism that it reduces the multiplicity of values to a common denominator, while utility theory in the hedonist era certainly did. Still, much confusion and pointless criticism would be avoided if locutions such as "welfare" or even "better off" were dropped. The second account of preference would be better phrased as saying that an agent prefers x to y if, in addition to the agent systematically selecting x over y, the agent also believes that there is greater justification for choosing x rather than y. For many purposes, the criterion of justification may be left as a black box.[3]

Rationality in ordinal preference theory is identified with two properties of $\succeq$: completeness and transitivity. A preference relation $\succeq$ is defined to be *complete* if, for all pairs of consumption bundles (x, y), either $x \succeq y$ or $y \succeq x$ (or both). An agent with complete preferences thus can at least weakly rank every pair of bundles. (The bundle x may be identical to the bundle y, and therefore complete preferences are always reflexive; that is, $x \succeq x$ for all x.) A preference relation $\succeq$ is defined to be *transitive* if, for all triples of consumption bundles (x, y, z), $x \succeq y$ and $y \succeq z$ imply $x \succeq z$. For the moment, think of transitivity as an internal consistency requirement. I will discuss rationales for completeness and transitivity in detail in the next section.

Like hedonism, ordinal preference theory employs utility functions, but it holds that their sole purpose is to summarize the information in preferences. A utility function is said to *represent* a preference relation if, for every pair of choices (x, y), the function reports that x has at least as much utility as y if and only if x is weakly preferred to y. In symbols, u represents $\succeq$ if, for every pair (x, y), $u(x) \geq u(y)$ if and only if $x \succeq y$. According to ordinal theory, a function that represents $\succeq$ is considered to be as good a summary of $\succeq$ as any other function that represents $\succeq$.

Ordinal utility functions, therefore, do nothing more than rank consumption bundles from best to worst. For you to grasp how limited this conception of utility is, let me simplify matters a little and assume that agents choose from only a finite number of different consumption bundles. (Real agents, of course, never have the chance to choose from sets that are any larger.) In this case, if $\succeq$ is complete and transitive, a utility function that represents $\succeq$ will always exist. Hence, the claim that agents maximize utility amounts to nothing more than an assertion that their preferences are complete and transitive. Specific functions that represent $\succeq$ can be assembled in a number of ways; perhaps the simplest is to let $u(x)$ equal the total number of options that x is strictly preferred to. So, for instance, a bundle x that is strictly preferred to none of the options is assigned utility 0, as are all options classified as indifferent to x.[4]

Ordinal utility functions are not cardinal (as defined in the preceding section). Beginning with a $u(x)$ that represents some preference relation $\succeq$, we could add a constant k, perhaps a very large number, to the utility of every bundle weakly preferred to some arbitrary bundle z. The new function, say $v(x)$, would thus assign the utility number $u(x) + k$ to all $x \succeq z$ and continue to assign $u(x)$ to the remaining x. The function $v(x)$ still represents $\succeq$ according to the ordinal definition of representation. But clearly, as long as z is not the least preferred bundle and there are at

least three bundles, $v(x)$ will not be an increasing linear transformation of $u(x)$; equivalently, some of the ratios of utility differences must change. In fact, the utility functions that represent $\succeq$ are precisely the set of increasing transformations of $u(x)$ (or, equivalently, the increasing transformations of $v(x)$) discussed in the previous section. Thus, unlike Jevonian theory and its additively separable utility functions, ordinalism does not suppose that cardinal yardsticks lie behind preference rankings.

Current official theory, therefore, substantially contracts the meaning of utility maximization. Ordinalists do not claim that agents form preferences by gauging how much utility their options deliver, or indeed that preference tracks any single psychological objective, much less a quasi-physical substance. Utility maximization means at most that agents' judgments about how to achieve well-being are complete—every pair of options is ranked—and that those judgments are transitively ordered, from which it follows (in the finite case) that the options can be put in a list from best to worst. Current economic theory thus makes more modest psychological claims than is often supposed. If there are difficulties in the economist's view of rationality, and I will argue that there are, they cannot be found in an allegiance to Benthamite or Jevonian psychology. Indeed, the concept of utility that economics now embraces is precisely the definition of utility, discussed at the end of section 2, that Jevons rejected as vacuous.

A Mathematical Note

If agents have complete and transitive preferences over a countably infinite set, utility functions representing those preferences will again always exist. If preferences are defined over an uncountable set of items, however, then complete and transitive preferences need not always be representable; there may be no function that assigns utility numbers to all potential items of choice that is consistent with the preference relation. But with an added technical condition—that there is a countable subset of items such that for each pair of items (x, y) with $x \succ y$, there exists a z in the subset that satisfies $x \succeq z \succeq y$—utility functions that represent $\succeq$ are again guaranteed to exist. (For details, see, e.g., Fishburn 1970 or Kreps 1988.)

5 COMPLETENESS *OR* TRANSITIVITY

At first glance, the ordinal preference model seems to be an unqualified improvement over its hedonist predecessor. By holding psychological

content to a minimum, preference theory rebuts the charge that it needs a reductionist account of human nature and avoids committing itself to a specific—and thus inevitably imperfect—psychology.

Ordinalism has also been able to lift the domain restrictions imposed by the first generation of utility theorists. Freed from the assumption that agents make decisions by weighing quantities of pleasure, ordinalists have happily extended preference theory to broader classes of decisions. Depending on the application at hand, agents are presumed to have rational preferences over abstract goods such as the absence of environmental degradation, over allocations that trade off material gain against ethical concerns, over the welfare of others, or even occasionally over political goals.[5] Moreover, since ordinalists take completeness and transitivity to characterize rationality per se, they deploy in these new domains the same axioms originally designed to model choice over material consumption goods. The removal of domain restrictions has opened even classically philosophical terrain to preference analysis; witness Harsanyi's (1953) claim that distributional equity should be determined by the decisions of rational agents who are ignorant of who in society they will ultimately be. Thus, the very topics that Jevons and others were reluctant to include in utility theory are now embraced by it.

But does ordinalism provide a convincing theory of rationality? More precisely, can it explain why a rational agent must obey the completeness and transitivity axioms? Hedonism, despite its implausibilities, did provide such an explanation. If each possible consumption experience can be placed on a single numerical scale of pleasure, any pair of consumption experiences can be compared and ranked—completeness is therefore satisfied. And since numbers are transitively ordered, the consumption experiences that generate these pleasure numbers are transitively ranked as well.

Of course, ordinalism is more general than hedonism in that hedonism provides just one way to justify completeness and transitivity. It may be possible to form preference judgments without carrying out a pleasure calculus, and as we will see presently, there are alternative rationales for transitivity as well. But because of hedonism's weaknesses, the formal generality of ordinalism does not by itself vindicate the ordinalist theory of rationality. Pleasure served an indispensable prescriptive function in early utility theory; agents who at first do not know how to choose between a pair of consumption bundles can resolve their impasse by investigating how much pleasure their potential choices deliver. In the absence of

a credible ordering principle, agents may not know how to rank their options. In formal terms, preferences can be incomplete: for some pairs of options x and y, agents may be unable to assert either $x \succeq y$ or $y \succeq x$. Ordinalism's open-mindedness about the motivations behind preferences, which is its main attraction, thus at the same time undermines its ability to justify one of its two key axioms.[6]

The difficulties caused by the lack of an ordering principle are less apparent in the case of traditional consumption goods; they stand out in the expanded domains that preference theory now tries to cover. Consider an agent trying to decide rationally how much of society's resources should be devoted to keeping the environment unspoiled. The agent acknowledges the force of several arguments: that both material wealth and keeping nature pristine are genuine goods, that nature should be treated with respect and even reverence, and that respect for nature does not entail that every glen should be preserved intact. Despite an awareness of these points—indeed, because of that awareness—the agent does not know where to draw the line in the conflict of ends. Recognizing the economic dimension to the problem, the agent approaches a specialist in the economic theory of rationality for help. The expert informs the agent, "You have a complete and transitive preference relation defined over ordered pairs of environmental cleanliness and material wealth. Choose a feasible ordered pair that is at least weakly preferred to all other feasible ordered pairs." The agent is at a loss; it was precisely in order to construct such preferences that the agent approached the specialist.

This story underscores a distinctive feature of rationality theory in economics: it does not take a stand on normative questions even though the agents it studies may well desire to have preferences that are normatively legitimate. This disengagement marks a clear departure from philosophical explanations of rationality that offer specific, substantive accounts of what is good, just, and legitimate. From this vantage point, the economic theory of rationality appears conspicuously incomplete.

Rational choice theorists will no doubt respond that the normative content of preference theory is limited to internal consistency; they are therefore excused from normative debates over substantive questions. But because of its agnosticism about motivation, ordinalism must concede that agents may want preferences that can be rationally defended. Such agents must deliberate about which normative criteria are appropriate and how they should be applied. To defenders of preference theory, this possibility presents no particular difficulty: they would claim that the

sources of preferences are not part of what the theory tries to explain. Hence, preference theory need not concern itself with, let alone resolve, deliberative predicaments. By this line of argument, it does not matter how normative questions ought to be resolved; it matters only that agents decide such questions one way or another. But the positive facts of preference are linked to normative theory since, as I mentioned, agents may desire preferences that are rationally defensible. Consequently, if agents are unsure about what is legitimate or substantively rational, their preferences may be ill-defined. Difficulties in normative theory thereby seep into the positive theory of preference.[7]

A standard challenge is put to any ostensible occurrences of incompleteness: force agents to choose. To find a preference between some pair x and y, inform agents that unless they choose one of the options, they will be assigned a third item that they are known to rank below both x and y. These forced choices are then identified as preferences. Since sufficiently dire threats can easily be devised, these elicitations will indeed generate an ordering of x and y. (If an agent responds that either x or y is acceptable, then both $x \succeq y$ and $y \succeq x$ are inferred, which, by definition, means that the agent is indifferent between x and y.)

Recall from section 4 that ordinal preferences have two main interpretations: they can refer either to agents' judgments of how best to promote their welfare or to choices. Evidently, our earlier argument for the incompleteness of ordinal preferences employed the welfare definition. The forcing procedure, in contrast, invokes the choice definition. Since this latter definition is, if anything, the dominant understanding of preference in economics, the forcing procedure presents a formidable case in favor of completeness.

But what of transitivity? Before scrutinizing forced choices on this score, let us return to the welfare interpretation of preference, and consider whether preference in this sense should satisfy transitivity. The welfare definition of preference is more demanding than the choice definition in that agents who think they will experience greater welfare with x than with y have a compelling reason to choose x over y. If they do not, they will end up with a worse outcome.[8] (In contrast, an agent who merely chooses x when y is available may be picking x only because of the need to make some choice.) The strongest arguments for transitivity cleverly exploit the fact that choices should track welfare judgments. Consider an agent who has well-defined welfare judgments over a set of three

alternatives. That is, assume that completeness is satisfied for each pair of items. If these preferences do not satisfy transitivity, we can label the options so that $x \succeq y$, $y \succeq z$, and $z \succ x$ hold. Suppose that option z is originally the status quo and we give the agent the opportunity to shift to y. Since the agent at least weakly prefers y to z, he or she will be amenable to the switch. Similarly, once y is the status quo, the agent should then agree to shift to x, which is strictly dispreferred to the original option z. Intransitivity can thus sequentially lead agents to inferior outcomes.[9]

A variant of this argument is the famous money-pump, originally due to Davidson, McKinsey, and Suppes (1955). Here agents exhibit a more blatant violation of transitivity: for some triple of options (x, y, z), preferences satisfy $x \succ y$, $y \succ z$, and $z \succ x$. Because each of these preferences is strict, such an agent, when originally endowed with z, will agree to part with a small amount of money to switch to y, then pay more money to switch from y to x, and then pay more money still to return to z, thereby ending up with the original status quo but with less money. If the judgments $x \succ y$, $y \succ z$, and $z \succ x$ are not altered by the loss of wealth, the agent can be subjected to more rounds of pumping.

The money pump has wielded remarkable influence. In its wake, even many critics of economic rationality have conceded that failures of transitivity will expose agents to a dire hazard. And the money pump does indeed provide grounds for why rational welfare judgments should satisfy transitivity. But, as I argued earlier, preference in the welfare sense is liable to be incomplete. Thus, any defense of the full ordinalist conception of rationality hinges on preference-as-choice and on whether preference in this sense, which is guaranteed to be complete, should be transitive as well.

Specifically, do the above sequential consistency arguments apply to choice? They may appear to apply. When preference is defined as choice, we may interpret the expression $a \succeq b$ to mean "Out of the set $\{a, b\}$, a is chosen" and $a \succ b$ to mean "Out of the set $\{a, b\}$, a is chosen and b is not." If we assume that at least one element is chosen out of every set—in accordance with the forcing procedure—then this preference-as-choice relation must be complete. Consequently, a violation of transitivity implies there is a triple (x, y, z) that satisfies $x \succeq y$, $y \succeq z$, and $z \succ x$. We now deploy the same sequence of exchanges used earlier: if z is the original status quo, the agent will agree to switch to y and then to x. Moreover, since $a \succeq b$ in effect means "The agent will accept a when b is

available," we do not need to worry at this point in the argument about any distinction between welfare and choice or about agents who agree to exchanges only when they have a strict welfare judgment.

To conclude that this sequence of choices is irrational, we must enrich the interpretation of $\succeq$ somewhat. As things stand, we have shown only that an intransitive chooser can end up with an option that is never directly chosen over the original status quo. If absolutely no welfare significance is imputed to $\succeq$, no irrationality can be inferred. But even a sliver of psychological content will bridge the gap. If we suppose that $a \succ b$ implies that the agent judges himself or herself to be better off with a than with b, then we may conclude that intransitive choosers are irrational: they end up with x even though they judge z to be superior. This interpretation of $\succeq$ is much less demanding than the ordinary welfare interpretation of $\succeq$, in which $a \succeq b$ implies that agents hold themselves to be at least as well off with a as with b. Here we impose interpretation only on agents' strict choices; that is, if agents *never* agree to accept b when a is available, we assume the agents are better off with a.

The above reasoning offers the strongest argument yet produced for equating rationality with the completeness and transitivity of preferences: interpret preference as choice and show that intransitive choices will expose agents to manipulation. Unfortunately, the manipulation conclusion hangs on a restrictive view of how agents must choose. In our interpretation of $\succeq$ as choice, we have assumed that $a \succeq b$ means that an agent will *always* choose a from the set $\{a, b\}$. But in the crucial case of agents who are unable to make welfare judgments over potential alternatives, that assumption is arbitrary and counterintuitive. Agents who cannot rank a pair of options a and b will sometimes choose a and sometimes choose b. Specifically, they may display *status quo bias*, in which they stick to the status quo until offered an alternative that they judge to make them better off. In our manipulation example, assume that y is unranked in welfare terms relative to both x and z, and in accord with the interpretation of strict choice given above, that z is ranked as superior to x. For concreteness, think of the alternatives as embodying different quantities of two rival goods and suppose that the agent is unable to rank trade-offs between the goods. The goods, for example, could be personal wealth for the agent and environmental quality, with x and z each containing more wealth but less environmental quality than y, and with z containing slightly more wealth and slightly more environmental quality

than x. Thus y would indeed be unranked relative to both x and z but z would be superior to x. How will an agent with these rankings choose? If z is the original status quo and the agent exhibits status quo bias, the agent, unable to judge how much wealth the environment is worth, will refuse to switch to y from z. Potentially manipulating sequences of exchanges thus never commence.[10]

It is crucial that status quo maintenance and other manipulation-avoidance strategies do not succeed by requiring that choice be transitive; otherwise, the traditional account of rational choice would be vindicated. To see the intransitivity of status quo maintenance, I need to introduce a new, less-restrictive definition of preference-as-choice. Observe that the forcing procedure, which I used to establish that preference-as-choice must satisfy completeness, by no means shows that agents must always choose the same element from any given set. The necessity of choice implies only that *some* option must be picked, and agents may want to vary their selections, perhaps to avoid manipulation or maybe out of whim. So let us instead interpret $a \succeq b$ to mean "There exist circumstances under which a is chosen from the set $\{a, b\}$," which is precisely what the forcing procedure demonstrates. Under this interpretation, the ability of agents to vary how they choose—say as a function of which option is the status quo—allows $\succeq$ to exhibit intransitivity while ensuring that agents are not manipulated. To confirm that intransitivity can occur in our example, note that although y is not chosen from the set $\{y, z\}$ when z is the status quo, it may well be chosen when y itself is the status quo. And similarly, x may well be chosen from the set $\{x, y\}$ when x is the status quo. We therefore have $y \succeq z$ and $x \succeq y$. Since the agent must always choose z from the set $\{x, z\}$ (z ranks strictly higher than x on welfare grounds), we have $z \succ x$ (that is, there are no circumstances under which x is chosen from $\{x, z\}$). Transitivity is therefore violated, and the case for the rational necessity of completeness and transitivity fails.

Status quo bias and other discordant evidence have been widely interpreted, by both economists and others, as a strong repudiation of the standard economic model of rationality. And status quo bias indeed contradicts the standard model. But, as I have indicated, the phenomenon is not a sign of irrationality in the sense that status quo bias puts agents in harm's way. Hence, it is not any thesis about the prevalence of genuinely rational behavior that must be overturned; it is the economic account of rationality that must give way.[11]

If we take a bird's eye view of the various arguments in favor of the ordinalist theory of rationality, a curious symmetry in their flaws appears. If preference is defined as a set of welfare judgments, then rational agents will satisfy transitivity but need not obey completeness; if preference is defined as choice, then although agents will definitionally satisfy completeness, rationality does not imply that they must obey transitivity.

The duality between preference-as-choice and preference-as-welfare-judgment illuminates some of the quarrels that perpetually beset preference theory. As I remarked, the domains nominally covered by preference analysis have steadily expanded. These expansions, moreover, have often been motivated by complaints that preexisting models of preference are psychologically too confining, that they do not allow agents' decisions to vary in a sufficiently rich way. This pattern of complaint and domain expansion is firmly established and will no doubt continue. For example, although there are exceptions, current economic models usually posit preferences that are defined over allocations of goods and not over the procedural rules that determine allocations. For instance, agents are typically assumed to care only about the decisions their government makes, not whether those decisions are determined by fiat or democratic vote. But if this tradition were subjected to sustained criticism, models would no doubt proliferate in which agents have preferences over allocations conjoined with procedural rules. Many critics protest that such conceptual moves leave preference theory vacuous and unfalsifiable. Defenders of orthodox preference theory, rarely persuaded by these charges, in turn reply that models with expanded domains do make falsifiable predictions. Transitivity, for example, is testable independently of the domain over which preferences are defined.[12] The present analysis points instead to a different drawback of mechanical domain expansions. Expansions occur when hitherto neglected aspects of decisions are incorporated into the definitions of the objects of choice. The new domains therefore usually describe a more complex class of decision problems. In the example above, for instance, agents would have to judge the equity and politics of various procedural rules and weigh those judgments against their attitudes towards allocations of goods. Incompleteness of preference is therefore far more likely, or if preference is defined as choice, intransitivity is more likely. The problem with domain expansions is not that they make preference theory unfalsifiable; rather, they render the ordinalist rationality axioms inapplicable.

6 PREFERENCE AS A SUBSTITUTE FOR ORDERING PRINCIPLES

I have illustrated the difficulty of constructing complete welfare orderings with the example of decisions that have a normative dimension; when agents want to do what is right, it is plain that they need a principle that shows them how to rank their options. But the incompleteness problem is not intrinsically tied to normativity. When agents have to choose between everyday consumption goods that deliver incommensurate but nonnormative benefits, they may not know which of their options best promote their well-being. Every inhabitant of the modern world is now and then defeated by the multiplicity of market choices. It is not just that we have too much information to process; the world of commodities simply cannot be reduced to a single ordering. Many decision quandaries are trivial—what flavor ice cream?—and have no abiding significance. But just as with momentous choices that pit the value of undisturbed nature against material wealth, the trivial dilemmas leave agents without a well-formed set of welfare judgments. And so people end up choosing in some other way. As I indicated in the previous section, these choices may end up displaying intransitivity, which in itself is evidence that agents' welfare judgments are incomplete (Raz 1986), but agents are not thereby exposed to the money pump or other hazards.

When available, ordering principles resolve the dilemmas of how to rank alternatives, often by showing that multidimensional decisions can be reduced to simpler choices over alternatives that agents already know how to order. Hedonism functioned in just this way in economics. It declared that seemingly complex consumption options, each of which combines apparently disparate and incomparable attributes, in reality all convey some quantity of a single sensation. The appeal of such a global ordering principle is manifest. In addition to the convenience of modeling agents with utility functions, pleasure ensures a determinacy to consumption decisions analogous to what profitability accomplishes in the business realm. The scale of profits, calculated in terms of money, provides firms with an external criterion that orders their production decisions objectively and unambiguously. In fact, many have considered the ready calculability of profits to be an essential cause of the dominance of means-ends calculations in modern societies. Hedonism extended the reach of instrumental rationality to cover all species of human decision making; each object of decision is made comparable in terms of its efficiency in delivering pleasure.

Hedonism in economics quickly came under rightful attack and had to be discarded. By substituting complete preferences for the ordering function previously performed by pleasure, ordinalism seemingly retained the advantages of utility maximization without its embarrassing psychological baggage. This strategy of replacing judgments about pleasure with preferences or choices followed the course set by the history of utilitarianism in moral philosophy. In Benthamite psychology, as it was commonly understood, all forms of desire—whether material wants, sympathy for others, or even a love of justice—are reduced to homogeneous pleasure. This reduction cannot be carried through, however, even for the desires of a single individual, and consequently Benthamism cannot guide preference and action. The young John Stuart Mill, for instance, complained that Bentham's philosophy was of little use to individuals deciding how to mold their "character" (Mill 1838). Of course, one can vacuously repair this incompleteness, though not its lack of prescriptive content, by declaring that the options that agents in the end choose are the ones that deliver the greatest pleasure. Mill himself took this tack in his later return to the utilitarian fold. Mill famously decomposed homogeneous Benthamite pleasure into qualitatively distinct types of pleasure (Mill 1861). How should one decide among the kinds of pleasure? Mill did not lay down any ordering principle; instead, one kind of pleasure is more valuable than another if those who are familiar with both prefer it. Moreover, in the cases he discussed, Mill claimed that the knowledgeable do in fact tend to choose as one. If this claim were correct, some prescriptive substance might be salvaged from Mill's position, but it is not.

Most, though by no means all, utilitarians in the twentieth century have followed Mill in rejecting the homogeneity of pleasure, in assigning primacy instead to agents' preferences, and in identifying whatever agents prefer as the more valuable pleasure or goal. The principal difficulty with this triad of moves is not the presumption that agents always opt for the more valuable goal. Since utilitarians typically place few restrictions on the ordering principles that agents may use to construct their preferences, this assumption need not impose a reductionist decision-making rule on agents. Indeed, when agents have a good grip on the comparative value of competing goals, and value is defined expansively, their preferences (and choices) will be guided by that understanding. This concordance between value and preference has no doubt bolstered the plausibility of preference-based utilitarianism. But the implication is only one-way: if agents cannot reach a firm conclusion about value, their choices obviously cannot reveal what they deem to be valuable. Like ordinal decision theory,

therefore, post-Benthamite utilitarianism lacks prescriptive content: it cannot guide preference or choice.

7 EXPECTED UTILITY THEORY: CARDINALITY REVISITED?

The theory of expected utility has long stood as the primary economic model of preference in the face of uncertainty. In the early days of neoclassical economics, Jevons and other pioneers offered little in the way of justification; they just asserted that agents maximize the mathematical expectation of their pleasure. For example, if $u(x)$ is the pleasure of option x and $u(y)$ is the pleasure of option y, the anticipated pleasure of receiving x with probability p and y with probability $(1-p)$ is given by the expected utility formula $pu(x) + (1-p)u(y)$. More generally, I assume in this section that there are a finite number of options, labeled $x^1, \ldots, x^n$. A typical prospect, often called a *lottery*, is denoted $(p^1, \ldots, p^n)$, where each p^i is the probability of receiving option x^i and where $\sum_{i=1}^{n} p^i = 1$. An agent who assigns the utility numbers $u(x^1), \ldots, u(x^n)$ to the n options therefore ascribes the pleasure level $\sum_{i=1}^{n} p^i u(x^i)$ to the lottery $(p^1, \ldots, p^n)$. (The superscripts in x^i and p^i serve as indices of the options and do not indicate that a quantity is raised to some power.) From our discussion of additive separability in section 3, it should be clear that expected utility functions are cardinal. That is, the functions u and v represent the same preferences over uncertain prospects (when each is inserted into the expected utility formula) if and only if v is an increasing linear transformation of u.

For Jevons and other early utility theorists, taking utility as a primitive fit nicely with their psychological views. But following the ordinalist revolution of the 1930s, utility could serve only as a tool to represent preferences and not as a theoretical starting point. With the raw material of the Jevonian approach missing, expected utility numbers could no longer be calculated. Conveniently, the mathematician John von Neumann and his coauthor Oscar Morgenstern soon accomplished the seemingly impossible, an axiomatization of the expected utility formula that takes ordinal preferences as primitive, even though expected utility functions are themselves cardinal. Like the ordinalist theory of section 4, the von Neumann–Morgenstern model begins with a preference relation $\succeq$ over a set of potential choices, but now that set consists of lotteries. Rationality is again identified with preference relations that satisfy the completeness and transitivity axioms. However, completeness and transitivity do not by themselves generate utility functions that satisfy the expected-utility for-

mula; two additional axioms are necessary. To explain these, I need to introduce compound lotteries, which are lotteries whose outcomes are themselves lotteries. For instance, a compound lottery might deliver lottery p with probability π and lottery q with probability $(1-\pi)$. Denote this lottery $(\pi p + (1-\pi)q)$. The von Neumann–Morgenstern theory supposes that agents regard a simple lottery as interchangeable with those compound lotteries that deliver the same final probabilities of outcomes; so $(\pi p + (1-\pi)q)$ is interchangeable with the simple lottery that delivers x^1 with probability $\pi p^1 + (1-\pi)q^1$, x^2 with probability $\pi p^2 + (1-\pi)q^2$, etc.

The first of the additional axioms, known as the *continuity* or *Archimedean* axiom, states that for every lottery r such that some lottery p is ranked strictly above r and some other lottery q is ranked strictly below r, there exists a lottery $(\pi p + (1-\pi)q)$ with $\pi > 0$ ranked strictly above r and another lottery $(\rho p + (1-\rho)q)$ with $\rho > 0$ ranked strictly below r. Continuity is so called because it presumes that if π is set near 1, then $(\pi p + (1-\pi)q)$ will be almost as desirable as p, while if ρ is set near 0, then $(\rho p + (1-\rho)q)$ will be almost as undesirable as q. The plausibility of the axiom hinges on whether or not the value of an outcome varies discontinuously as its probability changes. Although certainly not a feature of rationality per se, there obviously are many contexts in which agents will agree that their preferences should satisfy such a property.

The second and far more controversial additional axiom, *independence*, states that an agent weakly prefers lottery p to lottery q if and only if, for each lottery r and probability π, the agent also weakly prefers $(\pi p + (1-\pi)r)$ to $(\pi q + (1-\pi)r)$. In other words, if the agent prefers p to q, then he or she should still prefer p to q even after hearing the news that he or she might receive r rather than p or q. One argument in favor of the axiom goes as follows. Suppose that independence is violated, i.e., that for some p, q, and r, both $p \succeq q$ and $(\pi q + (1-\pi)r) \succ (\pi p + (1-\pi)r)$ hold. Now imagine that the agent has to choose between the compound lotteries $(\pi p + (1-\pi)r)$ and $(\pi q + (1-\pi)r)$. The choice proceeds in two stages. First, a coin that turns up heads with probability π and tails with probability $(1-\pi)$ is flipped. If tails, the agent receives r (and if r is a lottery, the remaining uncertainty about what option the agent receives is then resolved). If heads, the agent moves to stage 2, where he or she chooses between p and q. If the agent were to commit in advance to a choice at stage 2, the agent's options would represent the same alternatives as a one-stage lottery with options $(\pi p + (1-\pi)r)$ and $(\pi q + (1-\pi)r)$. It

seems reasonable, therefore, for the agent simply to plan to choose at stage 2 according to the dictates of $\succeq$. Given the preferences posited, the agent will plan to choose q. The coin is tossed. If tails, the agent receives r. But if heads, the agent, who by assumption has the preference $p \succeq q$, will want to choose p at stage 2, not q. The agent apparently cannot hold to preestablished plans, even in the absence of new information. (Remember: at the commitment stage, the agent knew that the choice between p and q would apply only if the coin were to come up heads.)

This inconsistency can readily be converted into a manipulation reminiscent of the money pump discussed in section 5. Suppose that an agent with the same preferences as above begins in possession of the compound lottery $(\pi p + (1-\pi)r)$. Since $(\pi q + (1-\pi)r)$ is strictly preferred to $(\pi p + (1-\pi)r)$, it is plausible that the agent will agree to switch from $(\pi p + (1-\pi)r)$ to $(\pi q^- + (1-\pi)r)$, where q^- has the same probabilities as q but each of the n options is now made slightly less attractive by subtracting a small amount of money from the agent's wealth. Suppose as before that the lottery $(\pi q^- + (1-\pi)r)$ proceeds sequentially. In stage 1, a coin is flipped that turns up heads with probability π and tails with probability $(1-\pi)$. If heads, the agent receives q^-, and if tails, the agent receives r. Stage 2 then resolves any remaining uncertainty in the lotteries q^- and r. The coin is now tossed. If tails, the agent receives r as planned. If heads, the agent is offered the chance to switch from q^- to p^-, a lottery with the same probabilities as p but with each of its n options diminished by a small amount of money. Since the agent regards p to be at least as good as q, he or she should strictly prefer p to q^- (by transitivity); hence if p^- is a small enough diminishment of p, the agent will prefer p^- to q^- and accept the offer. The agent has thus moved from an original position in which he or she receives r with probability $(1-\pi)$ and p otherwise to a position where he or she again receives r with probability $(1-\pi)$ but now receives p^- rather than p otherwise. The agent has traded away some expected wealth with no offsetting gain.[13]

This argument, known sometimes as "making book (or Dutch book) against oneself," has convinced many economists and decision theorists that independence is an inherent feature of rational conduct. But the Dutch-book argument relies on the implicit premise that $p \succeq q$ implies the agent ought also to prefer p to q (or p^- to q^-) *after* the coin toss. As Machina (1989) has argued convincingly, this premise is unwarranted. By supposition, the agent has the preference $(\pi q + (1-\pi)r) \succ (\pi p + (1-\pi)r)$. That is, when exposed to the possibility of receiving r with probability π,

the agent strictly prefers q to p. After hearing the news that he or she will not receive r, shouldn't the agent hold to this preference rather than to revert to the valuation that would have held had there never been a possibility of r? The fact that the agent did not receive r does not erase the earlier exposure to risk, and that exposure is as legitimate an influence on preference as past material consumption, which, according to all schools of preference theory, can properly affect current decision making. If the agent does treat past risk as equivalent to prospective risk, he or she will refuse the final switch to p^- and escape manipulation. This rebuttal does not completely settle matters—the relation $\succeq$ does not formally entail what preferences the agent will have after exposure to some uncertainty—but it weakens the case that violating independence necessarily invites manipulation.

But even if past exposure to risk can influence current preferences, there are certainly cases where individuals will concede that past risk ought not to bear on the present. One way to shift the persuasive ground in favor of independence is to change slightly the choice situation facing the agent with preferences $p \succeq q$ and $(\pi q + (1-\pi)r) \succ (\pi p + (1-\pi)r)$. Suppose that we present the agent with prospects p and q. The agent chooses p. We then announce that unbeknownst to the agent, there had earlier been a $(1-\pi)$ chance that the agent would have received r and not have been offered the choice between p and q. As it turned out, this eventuality did not materialize. Even though no barrier of logic or self-interest prohibits the agent from then reversing his or her choice between p and q, many would regard the news of the earlier possibility of r as irrelevant. To be free from Dutch-book manipulations, however, violators of independence must regard past and prospective risks as equivalent. Self-interest therefore requires that the preference between p and q be reversed. In circumstances where agents do not concur with the need for such reversals, the case for the rationality of independence gains ground.

Whether independence is intrinsic to rationality or not, a separate methodological consideration argues for applying normative preference theory only to decision-making problems in which independence can be expected to hold. To test the internal consistency of an agent's choices, we must observe several of the agent's decisions, and to ensure that these decisions are not spurious, there must be a chance that each choice determines which option the agent ultimately receives. The decisions therefore fall under the theory of choice under uncertainty. But if the agent can freely violate the independence axiom, almost any pattern of

choice will be consistent with virtually any rationality axiom. Suppose, for example, that we want to know whether an agent's preferences over the prospects p, q, and r are transitive. We present the agent in sequence with the choice sets $\{p, q\}$, $\{q, r\}$, and $\{p, r\}$, where each decision has the probability 1/3 of being the determining choice. If the agent were to choose p from the first set and q from the second, the agent—if he or she does not satisfy independence—may choose r from the third without violating transitivity. The first two choices, in fact, do not even reveal a preference for p over q or for q over r. The agent's objects of choice are triples of the form (p, q, r) that denote the decision made at each of the three choice sets; the agent therefore never makes a direct choice between p and q, between q and r, or between p and r. Indeed, without independence, we may infer only that one set of triples is preferred and that another set of triples is rejected. (I use "set" here because the agent may be willing to accept both items at one or more of the choice sets.) So, for example, if the agent were to choose (p, q, r), we would be able to infer only that the agent prefers (p, q, r) over (p, q, p), (p, r, r), (p, r, p), (q, q, r), (q, q, p), (q, r, r), and (q, r, p). Since the antecedent of transitivity—that some triple a is preferred to some triple b and that b is preferred to some triple c—does not obtain, testing transitivity is impossible. With independence, on the other hand, the preferences operative at one choice set must hold at the other two choice sets and overall. Axioms on preferences are then testable.

To sum up this lengthy digression, continuity and independence, even if they lack an ironclad claim to rationality, are certainly plausible in some circumstances. And independence is needed for empirical confirmation of any axiom on preferences.

When a preference relation $\succeq$ satisfies all four of the axioms we have discussed—completeness, transitivity, continuity, and independence—the von Neumann–Morgenstern expected utility theorem states that there exists a function u such that $p \succeq q$ if and only if $\sum_{i=1}^{n} p^i u(x^i) \geq \sum_{i=1}^{n} q^i u(x^i)$ (see, e.g., Fishburn 1970). How is it that the von Neumann–Morgenstern axioms on ordinal preferences generate a cardinal utility function? Various schools of preference theory have their answers to this question. Early on, some die-hard hedonists claimed that the von Neumann–Morgenstern theory resurrected the claim that utility is a measurable quantity. And some ordinalists have conceded that measurability of utility does obtain when the von Neumann–Morgenstern axioms are satisfied. But the majority position has held that the apparent cardi-

nality of von Neumann–Morgenstern preferences is a mathematical illusion (see, e.g., Arrow 1963, 10). While it is true that *within the expected utility formula*, u is unique up to an increasing linear transformation—that is, if $\sum_{i=1}^{n} p^i u(x^i)$ represents $\succeq$ and F is nonlinear, then $\sum_{i=1}^{n} p^i F(u(x^i))$ will not represent $\succeq$—we may apply a nonlinear transformation to the formula as a whole without changing the ranking of prospects. So if G is an increasing transformation, whether linear or not, then $G(\sum_{i=1}^{n} p^i u(x^i))$ must represent the same preference relation as $\sum_{i=1}^{n} p^i u(x^i)$. For example, with two options x^1 and x^2, $p^1 \log u(x^1) + p^2 \log u(x^2)$ does not represent the same preferences as $p^1 u(x^1) + p^2 u(x^2)$, since logarithms are nonlinear, but $\log(p^1 u(x^1) + p^2 u(x^2))$ does. It follows that agents who satisfy the von Neumann–Morgenstern axioms (or indeed any set of axioms on $\succeq$) need not experience well-defined ratios of differences in expected utility. And even when agents *can* report specific ratios of utility differences, agents with the same $\succeq$ can report ratios that differ.[14] Von Neumann–Morgenstern theory therefore does not present a genuine case of measurable utility.

The ordinalist consensus has used these arguments to try to remove any taint of measurability from the theory of choice under uncertainty. And it is certainly true that just as with choice over certain outcomes, agents do not need to assess uncertain options using a measurable concept of satisfaction or pleasure. Moreover, if we could be sure that preferences over uncertain options were complete, it would not matter for the theory how those preferences were formed. But how might agents go about assembling preferences over uncertain prospects? We saw in section 5 that simply compelling agents to choose will not generate preferences that satisfy transitivity. Agents must come to a reasoned judgment about how well the prospects available to them serve their interests. Yet it is not easy to gauge the value of lotteries. Agents must not only make judgments about the certain outcomes—say that x^1 is superior to x^2 and that x^2 is superior to x^3. To generate a complete ordering, they must also name a probability weighting of x^3 and x^1 that is indifferent to x^2. The obvious way to form such a judgment is to ask, "How does the gain from switching from x^2 to x^1 compare to the gain from switching from x^3 to x^2?" If we use the function u to gauge these "gains," the agent is in effect asking, "What is the following ratio?"

$$\frac{u(x^1) - u(x^2)}{u(x^2) - u(x^3)}$$

Not surprisingly, with this number in hand the needed probability is easy to calculate. (If we label the above ratio k, the agent will regard x^2 to be indifferent to the combination of x^1 with probability $1/(1+k)$ and x^3 with probability $k/(1+k)$.) If agents form preferences in this way, they are making explicit welfare judgments that single out the expected utility formula (or any of its increasing linear transformations) as psychologically accurate. Indeed, this method of building preferences amounts to a substantial return to Jevonian utility theory; agents have to *begin* with a utility or welfare judgment and derive their preferences from this primitive. A scalable sense of satisfaction would once again assume a pivotal prescriptive role in the theory of preference.

Assessing the cardinality of the von Neumann–Morgenstern construction is therefore delicate. By itself, the theory does not imply that agents form cardinality judgments or assess choices using a measurable gauge of satisfaction. But relative to choice under certainty, preferences over uncertain prospects constitute precisely the type of domain expansion that places the completeness axiom in doubt (see section 5). To qualify as an adequate account of uncertain choice, therefore, the von Neumann–Morgenstern approach must explain how agents come to rank the prospects they choose among. Conceivably, agents might assemble preferences without comparing measurable amounts of satisfaction. But until this missing psychological link is closed, cardinality remains the obvious device for forming preference judgments. Preference theory thus continues to rely on the Jevonian heritage it has worked so hard to jettison.

8 CONCLUSION: THE DOMAIN OF PREFERENCE ANALYSIS

Preference incompleteness diminishes the role that economic analysis can play in social decision making. Economics has long aimed to cut through the difficult debates that surround normative and political claims. Society does not need to resolve disputes over justice and right and the content of the good, it is said, since economic analysis can prescribe social reforms using only individual preferences as its raw material. While various economic theories of social choice do not agree on how individual preferences should be aggregated into social decisions, most camps agree that some principles are uncontroversial. For instance, virtually every normative theory in economics holds that Pareto-inefficient decisions should not be selected; that is, society should not adopt a policy if there is some other policy that leaves every agent at least as well off and that improves

the welfare of some agents. Such nostrums of policy advice presuppose that "at least as well off" and "improves the welfare of" are well-defined. These judgments about individual well-being are identified with individual preference orderings, which are, of course, assumed to be complete and transitive. This combination of taking preference as given and applying a mechanical aggregation procedure (such as Pareto efficiency) has given economic advice a technocratic air: a mechanical rule can generate correct, or at least better, social decisions. But if the presupposition of completeness is removed and the need for substantive ordering principles acknowledged, preferences no longer provide an adequate basis for policy analysis. Agents may not possess firm judgments about their own welfare, let alone about the well-being of the social whole, and their nascent judgments will be influenced by the very normative debates that economists have wanted to bypass.

These warnings do not mean that there are no domains to which preference analysis can be applied. In areas where agents do make welfare judgments—that is, where preference in the welfare sense satisfies the completeness axiom—rationality *does* require that choice satisfy the other classical rationality axiom, transitivity. The argument in section 5 that agents can choose intransitively without exposing themselves to manipulation applies only to options that are unranked on welfare grounds. Rational agents must always choose options that they think will deliver greater well-being; otherwise, they can end up with a worse outcome. The disjunction between choice and welfare judgment therefore occurs only when welfare judgments cannot be formed. Hence, when an ordering principle allows agents to come to definitive judgments about what promotes their well-being, the constraints of rationality are binding; if their choices do not satisfy transitivity, agents can be led to welfare-diminishing trades. And in those environments of uncertainty, discussed in section 7, where independence and continuity are plausible, the full machinery of expected utility maximization can be invoked.

The prerequisite of preference analysis, therefore, is to discover whether agents do have credible welfare rankings in the domain under study. Unfortunately, the standard vocabulary of economic theory, which simply equates choice and well-being, is poorly equipped to answer this question. Certainly, there are many cases of economic decision making in which agents find welfare rankings easy to construct. Decisions where income is the only interest at stake are clear cases; such decisions resemble the profitability orderings mentioned in section 6, which form the proto-

typical models of instrumental rationality. But just as certainly, there are numerous cases, particularly outside of economic life, where the completeness axiom is suspect at best. Harsanyi's model of social justice in terms of how individuals would choose if ignorant of what personality they will ultimately have provides a telling example. I began this essay by mentioning the expansion of economic preference analysis into other social sciences. No blanket assessment of this development is possible. The value of such analyses depends on where they fit on the spectrum between choice over monetizable goals and choice behind a veil of ignorance.

Acknowledgments

I am grateful for the detailed comments of Alyssa Bernstein, Elijah Millgram, and Jorge Restrepo.

Notes

1. Sen (1973, 1982, 1997) has long emphasized the differences between the choice and welfare definitions of preference. Levi (1986) also presses the distinction; his treatment of preference over uncertain prospects is particularly relevant to economic applications and to section 7 of this paper.

2. This result, whose history begins with Fisher (1892), requires that there are at least two goods k and j such that u_k and u_j are not constant functions. A technical restriction is also necessary, e.g., that the range of each u_i is an interval (as when the u_i are continuous). See Mandler (1999) for a relatively short proof and Krantz, Luce, Suppes, and Tversky (1971) for an extended treatment.

3. The persistent use of the word "welfare" in preference theory no doubt betrays a lingering belief that individuals *do* form their preference rankings by comparing welfare magnitudes. But many do not hold this view, and ordinalism in any event is not tied to this error.

4. It is easy to confirm that u represents $\succeq$. To show that if $x \succeq y$ then $u(x) \geq u(y)$, note that transitivity implies, for all z, that if z satisfies $y \succ z$, then $x \succ z$. The set of choices that x is strictly preferred to therefore contains the set of choices that y is strictly preferred to, and hence $u(x) \geq u(y)$. To show that if $u(x) \geq u(y)$ then $x \succeq y$, suppose to the contrary that $u(x) \geq u(y)$ and $y \succ x$. Reasoning as before, if z satisfies $x \succ z$, then $y \succ z$, and so the set of choices that y is strictly preferred to contains the set of choices that x is strictly preferred to. But the set of choices that y is strictly preferred to must have strictly more elements than the set of choices that x is strictly preferred to since, given $y \succ x$, x is in the former set but not the latter. Hence $u(y) > u(x)$, a contradiction.

5. The move to absorb ethical decision-making into preference analysis occurred early on, well prior to formal versions of ordinalism (see, e.g., Wicksteed 1910, book 2).

6. Many philosophical accounts of practical rationality offer alternative ordering principles, which could conceivably fill the prescriptive role once played by plea-

sure in utility theory. But until one of these accounts definitively explains how to justify preference, incompleteness is likely to persist: if agents are unaware of or unpersuaded by a proposed logic of decision-making, they will remain unable to order their alternatives.

7. The link between preference-as-it-is and preference-as-it-rationally-should-be is by itself compatible with orthodox preference analysis. As discussed in the introduction, that momentary behavior frequently violates rationality is widely conceded; rationality exerts its pull only gradually. The extra ingredient here is that there may be no identifiable set of welfare judgments that anchors preferences and that induces a well-defined ordering even in the long run.

8. "Welfare" is here defined as expansively as necessary, incorporating all due consideration for equity and the well-being of others. Still, even granting an expansive definition, there are cases that cloud the principle that rational agents should always choose the option that furthers their welfare. The act of choice may take on an independent meaning, as in Sen's (1973) famous example of the polite guest who desires the biggest slice of cake but chooses the second-biggest. Choice and the welfare definition of preference then need not coincide if the definition of welfare does not incorporate the symbolic import of choice. This problem disappears, however, if preference is defined as justified choice, following my suggestion in section 4.

9. If preferences obey continuity and nonsatiation conditions, the same conclusion will hold if agents agree to switch only to strictly preferred options (see Mandler 1998, where these conditions are defined precisely).

10. Agents can face more complex manipulations against which status quo bias does not provide adequate protection. For instance, consider an agent with the same welfare judgments over x, y, and z as in the example and who in addition strictly ranks a fourth option w strictly below x, y, or z. The agent faces a sequence of four choice sets, the first three of which are $\{y, w\}$, $\{y, z\}$, and $\{x, w\}$. The fourth set contains x and whichever option the agent chooses from the second set. To ensure that choice is not spurious, suppose that at each set the agent does not know whether there will be another round of choice. Status quo bias and optimization imply that y, y, and x, respectively, are chosen from the first three sets. The fourth set is therefore $\{x, y\}$. Since the agent selects x from the third set, status quo bias implies that x is also chosen from the fourth set. But, had the agent chosen z from the second set, he or she could then select z from the fourth set and be strictly better off (assuming the fourth set is the final round). Status quo bias therefore harms the agent relative to choosing according to a complete and transitive choice relation that agrees with the welfare judgments that the agent is capable of making. As I have shown elsewhere (Mandler 1998), however, more sophisticated choice procedures can immunize the agent from such manipulations and still exhibit intransitivity of choice.

11. Usually, evidence of status quo bias is taken as a sign that agents do not have preferences that are fixed through time (see, e.g., Tversky and Kahneman 1991). But although choice does indeed change through time depending on which option

is the status quo, there is no need to infer that welfare judgments also change. The choices that shift are likely to be between options that agents do not know how to order. Agents latch on to the status quo partly for this reason; when no ordering principle is apparent, the non-decision of holding to the status quo provides a convenient default decision.

12. Assuming, that is, that the domain is fixed during the time frame in which choice is observed. For example, to subject transitivity to empirical test, one cannot claim, after sequentially observing $x \succeq y$, $y \succeq z$, and $z \succ x$, that the x rejected in the third round is actually distinct from the x that is accepted in the first round. An independence assumption is also required; see section 7.

13. See Green 1987 for a more formal treatment and Machina 1989 for critical discussion. The antimanipulation rationale for independence was anticipated by early probabilists who argued that if agents' subjective probabilities do not conform to the rules of probability theory they will agree to bets under which they lose money with certainty (see de Finetti 1937 and Kyburg and Smokler 1964).

14. If G is nonlinear, and p, q, r, and s are four lotteries, the following ratios need not be equal:

$$\frac{\sum_{i=1}^{n} p^i u(x^i) - \sum_{i=1}^{n} q^i u(x^i)}{\sum_{i=1}^{n} r^i u(x^i) - \sum_{i=1}^{n} s^i u(x^i)}$$

$$\frac{G(\sum_{i=1}^{n} p^i u(x^i)) - G(\sum_{i=1}^{n} q^i u(x^i))}{G(\sum_{i=1}^{n} r^i u(x^i)) - G(\sum_{i=1}^{n} s^i u(x^i))}$$

The same $\succeq$ is therefore consistent with different ratios of utility differences.

References

Arrow, K. 1963. *Social Choice and Individual Values*. 2nd ed. New Haven: Yale University Press.

Davidson, D., J. McKinsey, and P. Suppes. 1955. "Outlines of a Formal Theory of Value, I." *Philosophy of Science* 22: 140–160.

De Finetti, B. 1937. "Foresight: Its Logical Laws, Its Subjective Sources." In *Annales de l'Institut Henri Poincaré* 7. Reprinted in *Studies in Subjective Probability*, edited by H. Kyburg and H. Smokler, pp. 93–158. New York: Wiley, 1964.

Fisher, I. 1892. *Mathematical Investigations in the Theory of Value and Price*. New Haven: Yale University Press, 1925.

Fishburn, P. 1970. *Utility Theory for Decision Making*. New York: Wiley.

Green, J. 1987. "'Making Book against Oneself,' the Independence Axiom, and Nonlinear Utility Theory." *Quarterly Journal of Economics* 102: 785–796.

Harsanyi, J. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 63: 434–435.

Jevons, W. S. 1871. *Theory of Political Economy*. London: Macmillan.

Krantz, D., R. D. Luce, P. Suppes, and A. Tversky, 1971. *Foundations of Measurement*. Vol. 1. New York: Academic Press.

Kreps, D. 1988. *Notes on the Theory of Choice*. Boulder: Westview.

Kyburg, H., and H. Smokler. 1964. Introduction. In *Studies in Subjective Probability*, edited by H. Kyburg and H. Smokler, pp. 1–15. New York: Wiley.

Levi, I. 1986. *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge: Cambridge University Press.

Lewin, S. 1996. "Economics and Psychology: Lessons for Our Own Day, from the Early Twentieth Century." *Journal of Economic Literature* 34: 1293–1323.

Machina, M. 1989. "Dynamic Consistency and Non-expected Utility Models of Choice under Uncertainty." *Journal of Economic Literature* 27: 1622–1668.

Mandler, M. 1998. "Incomplete Preferences and Rational Intransitivity of Choice." Harvard University.

Mandler, M. 1999. "Compromises between Cardinality and Ordinality, with an Application to the Convexity of Preferences." Mimeo, Royal Holloway College, University of London.

Marshall, A. 1920. *Principles of Economics*. 8th ed. London: Macmillan.

Mill, J. S. 1838. "Bentham." Reprinted in *Collected Works of John Stuart Mill*, edited by J. Robson, vol. 10, pp. 75–115. Toronto: Toronto University Press, 1969.

Mill, J. S. 1861. "Utilitarianism." Reprinted in *Collected Works of John Stuart Mill*, edited by J. Robson, vol. 10, pp. 203–259. Toronto: Toronto University Press, 1969.

Rabin, M. 1998. "Psychology and Economics." *Journal of Economic Literature* 36: 11–46.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Sen, A. 1973. "Behaviour and the Concept of Preference." *Economica* 40: 241–259.

Sen, A. 1982. Introduction. In his *Choice, Welfare, and Measurement*, pp. 1–38. Oxford: Blackwell.

Sen, A. 1997. "Maximization and the Act of Choice." *Econometrica* 65: 745–780.

Tversky, A., and D. Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics* 106: 1039–1061.

Wicksteed, P. 1910. *The Common Sense of Political Economy*. London: Macmillan.