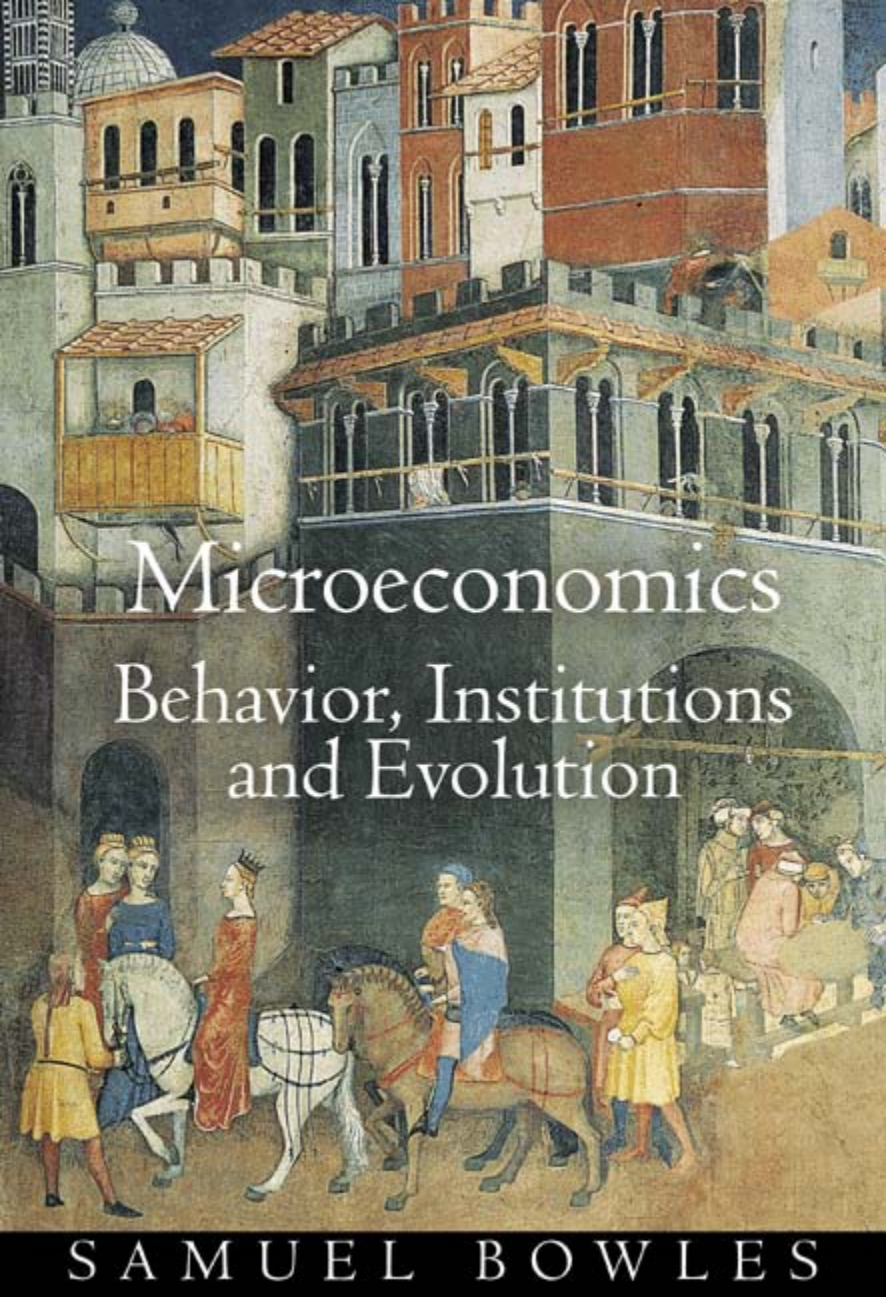
Microeconomics
Behavior, Institutions and Evolution
SAMUEL BOWLES

Microeconomics

The Roundtable Series in Behavioral Economics

The Roundtable Series in Behavioral Economics aims to advance research in the new interdisciplinary field of behavioral economics. Behavioral economics uses facts, models, and methods from neighboring sciences to establish descriptively accurate findings about human cognitive ability and social interaction and to explore the implications of these findings for economic behavior. The most fertile neighboring science in recent decades has been psychology, but sociology, anthropology, biology, and other fields can usefully influence economics as well. The Roundtable Series publishes books in economics that are deeply rooted in empirical findings or methods from one or more neighboring sciences and advance economics on its own terms—generating theoretical insights, making more accurate predictions of field phenomena, and suggesting better policy.

Colin Camerer and Ernst Fehr, editors

Behavioral Game Theory: Experiments in Strategic Interaction by Colin F. Camerer

Microeconomics: Behavior, Institutions, and Evolution by Samuel Bowles

Advances in Behavioral Economics, edited by Colin F. Camerer, George Loewenstein, and Matthew Rabin

The Behavioral Economics Roundtable

Henry Aaron
George Akerlof
Linda Babcock
Colin Camerer
Peter Diamond
Jon Elster
Ernst Fehr
Daniel Kahneman
David Laibson
George Loewenstein
Sendhil Mullainathan
Matthew Rabin
Thomas Schelling
Eldar Shafir
Robert Shiller
Cass Sunstein
Richard Thaler
Richard Zeckhauser

Microeconomics

BEHAVIOR, INSTITUTIONS, AND EVOLUTION

Samuel Bowles

RUSSELL SAGE FOUNDATION
NEW YORK
PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, 3 Market Place, Woodstock,
Oxfordshire OX20 1SY
And the Russell Sage Foundation, 112 East 64th Street, New York, New York 10021

Library of Congress Cataloging-in-Publication Data

Bowles, Samuel.
Microeconomics : behavior, institutions, and evolution / Samuel Bowles.
p. cm. — (The roundtable series in behavioral economics)
Includes bibliographical references and index.
ISBN 0-691-09163-3 (alk. paper)
1. Microeconomics. 2. Institutional economics. 3. Evolutionary economics.
I. Title. II. Series.

HB172.B67 2003
338.5—dc21 2003049841

British Library Cataloging-in-Publication Data is available

This book has been composed in Sabon

Printed on acid-free paper. ∞

www.pupress.princeton.edu

www.russellsage.org

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

List of Credits

Quotation from "Mending Wall" by Robert Frost used with permission of Henry Holt and Company.
Quotation from "16 tons" by Merle Travis used with permission of Warner-Chappell Music, a division of Warner Brothers.
Map of Italy in the fifteenth century (Figure 13.1) adapted from *Atlas of Medieval Europe* by Donald Matthew, used with permission of Andromeda, Oxford, Ltd.
Chapter 13 makes substantial use of work co-authored with Jung-Kyoo Choi and Astrid Hopfensitz that appeared in the *Journal of Theoretical Biology* (2003) 23:2, pp. 135–47, and is used here with permission from Elsevier.

For Libby and for Herb

Contents

CHAPTER THREE

Preferences and Behavior

> Political writers have established it as a maxim, that in contriving any system of government . . . every man ought to be supposed to be a *knave* and to have no other end, in all his actions, than his private interest. By this interest we must govern him, and, by means of it, make him, notwithstanding his insatiable avarice and ambition, cooperate to public good.
>
> —David Hume, *Essays*: *Moral, Political and Literary* (1742)

> Let us return again to the state of nature and consider men as if . . . sprung out of the earth, and suddenly, like mushrooms, come to full maturity without any kind of engagement to each other.
>
> —Thomas Hobbes *De Cive* (1651)

Growing corn is big business in Illinois. Using highly capital-intensive technologies and computer-generated business plans, some farmers cultivate a thousand or more acres, much of it on plots rented from multiple owners. In the mid-1990s, over half of the contracts between farmers and owners were sharecropping agreements, and over four-fifths of these contracts stipulated a fifty-fifty division of the crop between the two parties. In the southern part of the state where the soil is on average less fertile, there are counties where contracts giving the tenant two-thirds of the crop are common. In these counties there are few contracts of fifty-fifty or any division other than two-thirds, despite considerable variation in land quality within these counties.

Rice cultivation in West Bengal in the mid-1970s seems light years away from Illinois. Poor illiterate farmers in villages isolated by impassable roads much of the year, and lacking electronic communication, eked out a bare living on plots that average just two acres. We have already seen (in the Prologue) that they shared one similarity with Illinois's farmers, however: the division between sharecroppers and owners was fifty-fifty in over two-thirds of the contracts. (Ibn Battuta, whose visit to Bengal was also mentioned in the prologue, had noted—and deplored—exactly the same division of the crop six centuries before.) Other contracts were observed, but none of them constituted more than

The first epigraph is from Hume (1964:117–18), the second from Hobbes (1949:100).

8 percent of the total.[1] An even more striking example is from the U.S. South following the Civil War, where sharecropping contracts divided the harvest equally between the landlord and tenant irrespective of the quality of the land or whether the tenant was a freeborn white or a newly freed slave: "This form of tenancy was established everywhere in the South. It flourished with all possible combinations of soil quality and labor conditions"(Ransom and Sutch 1977:91, 215).

The puzzle of fifty-fifty sharecropping is the following: an equal split of the crop means that tenants on fertile land will have higher payoffs to their effort and other inputs than those on poor land. But if tenants are willing to work for the lower returns on the less good land, why should the owners of good land concede half of the crop to *their* tenants? The conventional economic theory of sharecropping predicts that the owner will capture the returns to land quality through variations in the crop share (Stiglitz 1974). But Burke and Young (2000) show that the Illinois sharecropping contracts allow the tenants on good land to capture a third of the differential return attributable to land quality, effectively transferring millions of dollars from owners to farmers.

A plausible interpretation of these facts is that farmers and owners around the world have hit on fifty-fifty as a seemingly fair division, and that attempts by owners to capture all of the returns to high quality land through the use of variable shares would be defeated by the tenants' retaliation. If true, this interpretation suggests that a predisposition to fairness, as well as the desire to punish those who violate local norms, may be motives as powerful as profit maximization and the pursuit of individual gain.

John Stuart Mill (1965[1848]) noted the striking global pattern of equal division in sharecropping, as well as local conformity to alternative shares in which fifty-fifty is not observed. Mill's explanation? "The custom of the country is the universal rule" (149). Custom may well be the proximate cause, but this explanation begs the question: why fifty-fifty as opposed to fifty-two–forty-eight? Why did the Bengalis and the Americans come up with the same number? We know from the analysis of the division game in chapter 1 that *any* exhaustive division of the crop is a Pareto-efficient Nash equilibrium: why this particular one? Even more puzzling: why does it persist when there appear to be huge profits to be made by offering lower shares on higher quality land? And when the shares do change, as we have seen happened in West Bengal in the 1980s and 1990s, why do they all change at once, reflecting the pattern of local homogeneity and punctuated equilibrium we encountered in chapter 2?

[1] Young and Burke (2001), Burke and Young (2000) and Bardhan (1984).

If motives such as fairness and retribution or simply adherence to convention override material self-interest in the highly competitive environment of Illinois agriculture it may be wise to reconsider the behavioral assumptions of economics, which conventionally has taken self-interest—summarized by the term *Homo economicus*—as its foundation. The need for a second look at *Homo economicus* is clear when considering problems of distribution such as sharecropping and other bargaining situations in which concerns with equity are likely to be salient. But the problem is much more general, and the canonical model of behavior seems to frequently fail even when fairness issues are absent.

Consider the following case (Gneezy and Rustichini 2000). Parents everywhere are sometimes late in picking up their children at day-care centers. In Haifa, at six randomly chosen centers, a fine was imposed for lateness (in a control group of centers, no fine was imposed). The expectation was that punctuality would improve. But parents responded to the fine by even greater tardiness: the fraction picking up their kids late more than doubled. Even more striking was the fact that when after sixteen weeks the fine was revoked, their enhanced tardiness persisted, showing no tendency to return to the status quo ante. Over the entire twenty weeks of the experiment, there were no changes in the degree of lateness at the day-care centers in the control group.

The authors of the study reason that the fine was a contextual cue, unintentionally providing information about the appropriate behavior. The effect was to convert lateness from the violation of an *obligation* that the parents were at some pains to respect, to a commodity with a price that many were willing to pay. They titled their study "A Fine is a Price" and concluded that imposing a fine labeled the interaction as a market-like situation, one in which parents were more than willing to buy lateness. Revoking the fine did not restore the initial framing of punctuality as an obligation, it just lowered the price of lateness to zero. The fact that monetary incentives for punctuality induced even greater tardiness is both counter to the predictions of the standard economic model and of general relevance to the problem of designing effective contracts and economic policies. In Hume's terms, the Haifa day-care centers designed a constitution for knaves, and they seemingly produced knaves rather than improved behaviors.

The weaknesses of the conventional model suggested by the puzzle of the fifty-fifty crop share and the fact that fining the Haifa parents backfired are evident in arenas of more conventional economic interest, such as labor markets, team production, tax compliance, the protection of local environmental commons, and other forms of public goods provision. Included is the importance of fairness motives in wage setting and other exchanges (Bewley 1995, Blinder and Choi 1990). Equally puz-

zling in the standard paradigm is the fact that individuals bother to vote given that the likelihood that their vote is decisive is vanishingly small, as well as their significant support, when they do vote, for tax-supported income transfers to the poor even among those sufficiently rich and upwardly mobile to be very unlikely ever to benefit directly from them (Fong 2001, Gilens 1999). Finally, studies at Continental Airlines, Nucor Steel, and other companies have found that group incentives are effective even where the gains are shared among such a large number that the individual payoff to one's own effort is negligible (Hansen 1997, Knez and Simester 2001).

Seeking a more adequate behavioral foundation for economics and the other social sciences, in this chapter I draw upon recent research to present a reformulation of the standard approach, one that retains a central role for individuals' preferences, beliefs, and constraints in explaining what people do, while emending the conventional model in three ways.

First, many behaviors are best explained by what are termed *social preferences*: in choosing to act, individuals commonly take account not only of the consequences of their actions for themselves but for others as well. Moreover they often care not only about consequences but also about the intentions of other actors. An important example of social preferences are *reciprocity* motives, according to which people are generous toward those who have behaved well (toward them or others) while punishing those who have not behaved well. Reciprocity motives induce people to act this way even in situations (such as one-shot interactions) in which generosity and punishing behaviors are personally costly and bear no expectation of subsequent or indirect reward. (These cases are examples of what I term *strong reciprocity*, to distinguish this behavior from reciprocation with the expectation of future reward, sometimes termed *reciprocal altruism*.) Other social preferences to be considered are *inequality aversion*, *envy* (or spite), and *altruism*.

By contrast, the conventional assumption is that individual behavior is entirely explained by what is loosely termed self-interest, by which I mean *self-regarding preferences defined over outcomes*. According to this view, our concerns extend neither to the outcomes experienced by others nor the processes generating the outcomes. F. Y. Edgeworth, a founder of the neoclassical paradigm, expressed this view in his *Mathematical Psychics* (Edgeworth 1881:104): "The first principle of economics is that every agent is actuated only by self-interest." Self-interest is not presumed by rationality (one could have transitive and complete altruistic or masochistic preferences), but it is commonly treated as axiomatic in economics (and sometimes confused with rationality). Thus, while self-interest is not formally implied by the conventional approach, it is generally assumed in practice. The assumption acquires consider-

able predictive power in strategic situations when it takes the form of what I term the *self-interest axiom*, namely, individual self-interest coupled with the belief that others are also motivated by self-interest.

Second, individuals are *rule-following adaptive agents*. By this I mean that we economize on our limited cognitive resources by acting according to evolved rules of thumb. The term "boundedly rational" is sometimes used to describe the cognitive limits of real human actors, but I do not use it as it suggests irrationality. It is not the boundedness of our rationality that I would like to stress but rather our limited capacity and predisposition to engage in extraordinarily complex and costly cognitive exercises. Among these evolved behavioral rules are ethical prescriptions governing actions toward others, namely, *social norms*, conformity to which is both valued by the actor (i.e., the norm is internalized) and supported by social sanction. This approach contrasts with the conventional view in which behavior is the result of often quite demanding individual cognitive processes addressing both evaluative and causal issues (is this state desirable? how can I bring it about?). This conventional *individual cognition-centered* view excludes behavior based on such things as visceral reactions (like disgust, fear, or weakness of will), habit, or evolved rules of thumb, and it presumes (against a considerable body of evidence) that individuals are both able and predisposed to make quite advanced inferences about what others will do and about the way the world works.

Third, behaviors are context dependent, in three senses. Situational cues are used to determine the behaviors appropriate in any given setting. Moreover, we evaluate outcomes from a particular point of view, namely, our current state or the state experienced by a member of our reference group. Finally, social institutions influence who we meet, to do what, and with what rewards; as a result, our motivations are shaped through the process of cultural or genetic transmission introduced in chapter 2. Thus, our *preferences are situationally specific and endogenous*. If one's experiences result in durable changes in preferences, they are said to be endogenous, which will happen if experiences affect either social learning or (over the very long run) genetic inheritance. This may be compared with situation- or state-dependent preferences that are time invariant (over time, one behaves the same way in the same situation). Because endogenous preferences involve learning or genetic changes, behavior in the same situation changes over time.

This approach contrasts with the conventional view that preferences do not depend on one's current state and are either unchanging or change solely under the influence of influences exogenous to the problem under investigation. George Stigler and Gary Becker (1977) expressed this view in their essay *De Gustibus Non Est Disputandum*: "One does not argue about tastes for the same reason that one does not

argue about the Rocky Mountains—both are there, and will be there next year, too, and are the same to all men" (76). They were repeating, in less poetic terms, Hobbes' point about mushrooms.

Nobody takes the conventional assumptions literally, of course. Edgeworth observed that the self-interest assumption is literally true only in limiting situations ("contract and war"), and Hume, in the sentence immediately following this chapter's first epigraph, mused that it is "strange that a maxim should be true in politics which is false in fact." Hobbes invoked a deliberately fanciful analogy to abstract from the social formation of preferences as part of a thought experiment, not as a description of real people.

While recognizing that the standard assumptions are often violated empirically, most economists have shared Becker and Stigler's endorsement of the simple canonical model of exogenous and self-regarding preferences. The broad acceptance of its tenets—not as empirical truths but as close enough approximations to be useful analytical shortcuts—is explained in part by their substantial contribution to both intellectual discipline and clarity. The standard assumptions provide a common intellectual framework resistant to ad hoc explanation on the basis of empirically unobserved individual differences or changes in tastes over time. Abandoning the standard model opens the door to explanations of behaviors on the basis of vague concepts like "psychic income" or "animal spirits."

For a new behavioral foundation to be a contribution to social science rather than an invitation to ad hoc explanation, we need more empirical information about preferences and how we come to have them as well as more adequate models of behavior under less restrictive preference assumptions. The extraordinary production of empirical findings by experimental and behavioral economists and other social scientists in recent years has made such a reformulation not only possible but overdue. Here and in later chapters, I make extensive use of experimental results. The reason is that this relatively new method in economics has for the first time allowed the testing in controlled settings of well-formulated hypotheses concerning the behavioral assumptions of economics.

In the next section I introduce what I call a behavioral interpretation of preferences and rational action, followed by a review of a number of empirical anomalies in the conventional treatment of preferences. I then turn to recent research on social preferences, introducing both experimental results and two new utility functions. I postpone until chapters 7, 11, and 13 the formal modeling of how preferences evolve, why people often adhere to ethical norms, and why other-regarding motives such as generosity and fairness are common.

Preferences, Reasons, and Behaviors

When individuals act, they are generally trying to *do* something, however wisely or otherwise. An implication is that individuals' purposes and their understandings about how to carry them out, along with the constraints and incentives posed by societal rules and individual capacities, are key ingredients in accounting for individual actions. What people do in any situation therefore depends on their preferences and their beliefs.

Beliefs are an individual's understandings of the relationship between an action and an outcome. In many cases beliefs enter trivially in choice situations and so are not explicitly addressed: we routinely assume for example that people know the payoff consequences of their actions in simple games. In other situations—particularly in strategic interactions without dominant strategies—beliefs may become all important: the effect of my attending a meeting may depend on who else is there and so my decision to attend or not will depend on my expectation of who else will attend, which in turn will depend on *their* beliefs about whether others will attend, and so on. In other situations the structure of the interaction may be ambiguous and understood differently by different players. In these situations, how we come to have the beliefs we do and how we update our beliefs in light of our experience assumes central importance.

Preferences are reasons for behavior, that is, attributes of individuals—other than beliefs and capacities—that account for the actions they take in a given situation.[2] Preferences thus include a heterogeneous melange: tastes (food likes and dislikes, for example), habits, emotions (such as shame or anger) and other visceral reactions (such as fear), the manner in which individuals construe situations (or, more narrowly, the way they frame a decision), commitments (like promises), socially enforced norms, psychological propensities (for aggression, extroversion, and the like), and one's affective relationships with others. To say that persons act on their preferences means only that knowledge of the preferences would be helpful in providing a convincing account of the actions (though not necessarily the account that would be given by the actor, for as is well known, individuals are sometimes unable or unwilling to provide such an account).[3]

This "reasons for behavior" interpretation of preferences may be con-

[2] A more precise term for this conception of preferences might be the cumbersome expression suggested by Nowell-Smith (1954): "pro and con attitudes."

[3] See Nisbett and Wilson (1977). Shafir, Simonson, and Tversky (2000) provide an interpretation of what they call "reason-based choice" similar to that in Nowell-Smith and here.

trasted with two conventional approaches. The first postulates that individuals seek to maximize their utility, equating utility to well-being, pleasure, or happiness, in the tradition of Jeremy Bentham and the early nineteenth-century utilitarians. In the more recent revealed preference approach, by contrast, a preference ordering is nothing more than a complete *description* of consistent behavior, and any connection to a hedonistic calculus is gratuitous. Neither approach is entirely adequate.

If our objective is to explain behavior, the revealed preference approach is vacuous because it is silent on the question of motives and reasons: while these are hardly sufficient to an explanation, they are rarely uninformative. The revealed preference view once attracted adherents impressed by the now-antiquated methodological fiat that subjective states are not knowable, so a scientific approach must focus on observable behaviors. By contrast, the utilitarian approach is substantive; the subjective states central to this view—pleasure, pain, satisfaction, anxiety, and other hedonic experiences—are now an active field of scientific study and measurement. But treating behavior as synonymous with the pursuit of well-being is misleading: the reasons for our actions also include addictions, weakness of will, myopia, and other well-documented dysfunctional aspects of human behavior. The fact that the same term—utility—is conventionally used both as an explanation of behavior and as a standard for evaluating social outcomes has forced economists to take an unduly limited view of both behavior and social evaluation.

To review thus far, along with the set of feasible actions and the associated outcomes, beliefs and preferences provide an account of individual action. Recall that I have defined institutions as the population-level laws, informal rules, and conventions that give a durable structure to social interactions. In game theoretic terms, an institution is a game (which, as we have seen in chapter 1, may also be the outcome of an underlying game), preferences are the evaluation of the payoffs, and beliefs are the players' understandings of the expected payoff consequences of each strategy in their strategy set (i.e., their understanding of the game and its payoff structure plus the likelihood of others' actions).

As preferences, beliefs, and institutions are easily confounded, consider a concrete case. The common practice in many countries of driving on the right-hand side of the road is an institution; it is a convention, that is, an equilibrium of an Assurance Game, and the convention is supported by laws. In these countries it is a best response to drive on the right, and it is also illegal to do otherwise. People do not *prefer* driving on the right, per se, they prefer avoiding crashes and fines, and were everyone else to drive on the left without breaking the law, they would drive on the left as well. The belief that others will drive on the

right sustains the institution of driving on the right, which in turn sustains the belief. Beliefs and preferences are facts about individuals that sustain this particular equilibrium, while institutions—represented in this case by the driving-on-the-right equilibrium—are facts about groups of people.

A version of the beliefs and preferences framework, which I will term "conventional," has provided the behavioral foundation for economics and is increasingly applied throughout the social sciences. An individual's behavior is modeled using a utility function: $U = U(x, y, z)$. The arguments of U—x, y, and z—describe a *state* that may be a simple list of goods consumed or more complex formulations like a cold beer on a hot evening three days from now in the company of friends in an Islamic society that prohibits the consumption of alcohol. The utility function is taken to be a numerical representation such that higher values of U are chosen (said to be preferred) over lower values, the state (x, y, z) being chosen over (x', y, z) if U $(x, y, z) > U(x', y, z)$.

The utility function is *complete*, meaning that every state can be ordered by a relationship of either preference or indifference with respect to every other state. The ordering is also *transitive*, meaning that the orderings it gives do not include inconsistent orderings such as (x, y, z) preferred to (x', y, z), which is preferred to (x'', y, z), but (x'', y, z) is preferred to (x, y, z). Finally the utility function is (usually implicitly) assumed to be *time invariant* over the relevant period: when, say, prices change exogenously, the individual responds to the new prices and not also to coincident changes in the utility function. When individuals act according to a complete and transitive utility function they are said to be *rational*.[4] Other ways of acting—inconsistency of choice induced by whim or incompleteness of preferences over unimaginably horrible outcomes, for example—are not thereby deemed *ir*rational, of course, they are simply forms of action not covered by this model perhaps better deemed *non*rational.

The conventional model is routinely extended to cover risk and uncertainty. *Risk* is said to exist if a consequence of an action in the individual's choice set is a set of possible outcomes each occurring with a *known* probability. By contrast, if one or more of the actions open to the individual may cause more than one outcome, the probabilities of which are *unknown*, *uncertainty* exists. Both are ubiquitous aspects of choice. Deciding whether to rent a cottage at the beach knowing that with probability p it will rain is an example of risk. In these cases the

[4] Other rationality restrictions are sometimes imposed. For example, the weak axiom of revealed preference requires that if (x, y, z) is preferred to (x', y, z) then (x, y, z, a) will be preferred to (x', y, z, a).

individual is assumed to maximize *expected utility*. The expected utility of an action is the utility associated with each possible consequence of the action multiplied by the probability of its occurrence: U(beach cottage) = $(1 - p)U$(beach cottage in the sun) + pU(beach cottage in the rain).

The maximization of expected utility requires more than the simple ordering of each possible state (that suffices to determine behavior under certainty) as it uses information about how much better one state is than another. In a pioneering work on game theory, John von Neumann and Oskar Morgenstern (Neumann and Morgenstern 1944), showed that an expected utility maximizing individual's choices are invariant for additive or linear transformations of the utility function. (What this means is that if an individual's behavior is described by the utility function u then her behavior is also described by any function of the form $v = \alpha + \beta u$ where $\beta > 0$.) What are termed *von Neumann-Morgenstern utilities* embody this restriction. They have already made two unannounced appearances in chapter 1: in the treatment of risk dominance, and when I normalized the payoffs associated with the fallback positions in the conflict of interest games. Von Neumann-Morgenstern utilities exhibit cardinality over the states *for a given individual* but not *between* individuals; they indicate how much better the beach in the sun is compared to the beach in the rain *for you*, but not how much better either is *for you* than *for me*. All of the payoffs subsequently used here are Von Neumann-Morgenstern utilities unless specified otherwise.

In the case of uncertainty, the known probability weights are replaced by the individual's subjective estimates of the unknown probabilities. It is generally assumed that individuals modify their estimates on the basis of recent experience by a process termed *Bayesian updating*; Reverend Thomas Bayes (1702–1761) was an early writer on probability theory. The Bayesian approach to rational action assumes that individual decision making under uncertainty is based on expected utility maximization based on subjective probabilities updated in this manner. (The Bayesian approach obviously presumes von Neumann-Morgenstern utilities.) The difference between risk and uncertainty in practice is often blurred except in limiting cases, where truly known probabilities are involved such as allocation mechanisms that are randomized by a coin toss.

An important application of these ideas is the concept of *risk aversion*, measured by the degree of concavity of a utility function $U(W)$, where W is the wealth of the individual. The intuition is that if the marginal utility of wealth is sharply declining in wealth, as will be the case for a "very concave" utility function, then one would value $75,000 with certainty a lot more than an even chance of $50,000 or

$100,000. Thus, an individual whose utility is concave in wealth will be averse to a lottery over two prizes if she could have, instead, a certain prize equal to the expected value of the lottery. For this reason, a measure of the degree of risk aversion is $-U''/U'$, called the Arrow-Pratt measure.[5] An individual is *risk neutral* if utility is linear in wealth or $U'' = 0$; $U'' > 0$ implies *risk seeking*.

A second essential extension is to choices over states at different dates. This is accomplished by discounting future states at a constant *discount factor* δ, which is an inverse measure of the degree to which we discount future events due to myopia, the unlikelihood of surviving to some future date, and other reasons.[6] For a person who values future states the same as current states, $\delta = 1$ while for more present oriented individuals, $\delta < 1$. According to the *discounted utility* approach, δ is defined such that an individual is indifferent between adding x to her consumption y at time t and adding some other increment, x', n periods later, at $t + n$ if

$$U(y + x)\delta^t + U(y)\delta^{t+n} = U(y)\delta^t + U(y + x')\,\delta^{t+n} \qquad (3.1)$$

Thus, extended to cover risk and intertemporal choice, the conventional model captures the important intentional aspect of human behavior and combines broad applicability with formal tractability. At first glace it appears to impose few substantive restrictions on the analysis of behavior other than the exclusion of the perhaps unimportant cases of incompleteness and inconsistency just mentioned. But this is not correct: the above formulation is a substantive theory of behavior, and embodies strong claims about what kinds of things people take account of and how they do this. This model does not fare well in light of recent empirical research about behavior.

Situation-Dependent Preferences

One of the best documented falsifications of the conventional model arises because preferences (and hence behaviors) are *situation dependent* in the following sense. Suppose ω_i is a vector representing a state i (e.g., one described by (x,y,z) above), an element of the set of possible states Ω, and $U_i(\omega_j)$ is the utility associated with state $\omega_j \in \Omega$ *for an individual currently experiencing state* ω_i. Let $U_i(\omega)$ represent this individual's preference ranking of all the possible states when that individual is in state i. Then preferences are situation dependent if the rankings

[5] See Mas-Colell, Whinston, and Green (1995) for further elaboration.

[6] The discount factor $\delta = 1/(1 + r)$ where r is the rate of time preference.

by the same individual in a different state, given by $U_k(\omega)$ differ from those given by $U_i(\omega)$ for some i and k. Situation dependence is also called state dependence, but I use the former in recognition of the substantial literature in psychology on the importance of situations as influences on behavior.

An important example of situation dependence, termed *loss aversion*, arises because people value losses (negatively) more highly than equivalent gains. The size of the loss aversion coefficient is surprisingly large: estimates from both experiments and natural settings find that the disutility of a small loss is between two and two-and-a-half times the utility of a small gain. The utility function is thus sharply kinked at the status quo (and the kink moves when the status quo changes). Closely associated is the *endowment effect*: the minimal price that would induce an individual to sell something she now possesses is substantially higher than the maximum price she would be willing to pay to acquire the same good. (Loss aversion and the endowment effect are examples of a broader class of situation-dependent effects, namely *status quo bias*.)

Loss aversion and endowment effects have been extensively documented in experiments by economists and psychologists, and they provide plausible explanations of important anomalies in everyday economics. For example, the fact that U.S. stock returns have consistently exceeded bond returns by a wide margin is an outstanding puzzle in economics. It was once thought to be a result of risk aversion among investors, but a simple calculation (Mehra and Prescott 1988) shows that the level of risk aversion necessary to explain the difference is implausibly large. For risk aversion to account for the stock return puzzle, investors would be indifferent between an even chance of $50,000 and $100,000 and a sure thing of $51,209. A more compelling account (Benartzi and Thaler 1995) holds that investors are not averse to the variability of returns per se (after all, most are quite rich), but they react strongly to the prospect of losses, and stock returns over a year are negative much more often than bond returns.

The loss aversion interpretation of the stock return puzzle makes it clear that a precise formulation of loss aversion and other aspects of situation-dependence requires explicit treatment of the time dimension; if investors had a five-year time horizon, they would experience few negative returns, so the loss aversion explanation implies a particular time horizon, evidently a rather short one. An individual who experiences a loss will eventually treat the new situation as the status quo. We know, for example, that people who anticipated that a severe physical handicap would be unbearable often become quite satisfied with life after living with the handicap for a matter of years. A well-documented situational determinant of preferences is simple exposure (Zajonc 1968). People come to value more the things (for example, foods) they've been

exposed to longer. Rats are no different: those brought up on Mozart prefer his music to Schoenberg (Cross, Halcomb, and Matter 1967). Sometimes preferences adjust to situations virtually instantaneously—students in endowment-effect experiments bonded with the coffee mugs given them in a matter of minutes!—but the lags are considerably greater in many cases.

Situation dependence—in the form of loss aversion, endowment effects, and long-term endogeneity of preferences—by no means exhausts the empirical shortcomings of the conventional model. Like the assumption of situation independence, the conventional treatment of intertemporal choice is strikingly counterintuitive and strongly contradicted by behavioral evidence.[7] Suppose you were indifferent between one meal at your favorite restaurant now and two such meals a year from now. Then according to eq. (3.1) you would also be indifferent between one meal (call it x) twenty years from now and two meals (that's x') twenty-one years from now. To see this, notice that this indifference relationship can be equivalently expressed (divide both sides of (3.1) by δ^t) as

$$U(y + x) - U(y) = \{U(y + x') - U(y)\}\delta^n.$$

Thus the difference in your utility made by the delay of the two meals does not depend on when it happens in real time, but only on the amount of time elapsed between the time of the first (one-meal) and the second (two-meal) event. This so called *stationarity property* of the discounted utility model is a temporal analogue to state independence: *how one evaluates states is assumed not to depend on where one is evaluating them from*. This is not only counterintuitive; it is contradicted by extensive experimental and other evidence (interestingly, for other animals as well as humans). For most people, as the example suggests, the delay of a year is a lot more salient if it occurs sooner rather than later, suggesting what is called a *hyperbolic discount function*, according to which a state in year t is discounted not at the rate δ^t but instead at the rate

$$\delta(t) = (1 + \alpha t)^{-\beta/\alpha} \qquad \text{with} \qquad \alpha, \beta > 0 \tag{3.2}$$

which for large values of α indicates that the value of future states is rapidly declining in the near future, after which the decline is sharply attenuated (so that, for example, you might be quite impatient about waiting a year for your favorite meal but only somewhat less impatient in evaluating the long-term consequences of global warming).[8] Hyperbolic discounters will exhibit preference reversal behavior: of two prizes

[7] This paragraph draws on Loewenstein and Prelec (2000).

[8] The departure from constant discounting is governed by α; you may confirm that as α goes to zero eq. (3.2) reproduces the standard exponential discount function $\delta(t) = e^{-\beta t}$.

A and *B* of differing amounts and occurring at different future dates, one may prefer *A* over *B* at the present but with the passage of time prefer *B* over *A*. A hyperbolic discounter might, for example, take the one meal now over the two meals a year from now but also choose the two meals twenty-one years from now over the one meal twenty years from now. But if this is the case, after the passage of nineteen years, the hyperbolic discounter would choose the one meal sooner over the two meals later, thus reversing his choice. A number of studies (surveyed in Angeletos, Laibson, Repetto, Tobacman, and Weinberg 2001) suggest that the hypberbolic discounting approach provides better predictions than the conventional approach of individual savings behavior, accounting for the empirically-observed significant increases in consumption from predictable increases in income, and the sharp reduction in consumption upon retirement.

As in the case of intertemporal choice, well-established empirical regularities are anomalous from the standpoint of the conventional expected utility analysis of choice in the presence of risk. Recall that this framework requires that individuals evaluate the actions they may take according to the linear sum of the probability of each possible consequence occurring, multiplied by the utilities associated with each consequence. Thus, events occurring with arbitrarily small probability should be treated virtually indistinguishably from events that will certainly not occur. But it is well established that people do not evaluate lotteries over risky events this way: an event that will happen with certainty is regarded as quite different than something that will happen with probability $(1 - \varepsilon)$, no matter how small ε is. Conversely, knowing that one is not HIV positive is hardly the same thing as knowing that one may be HIV positive, but with an arbitrarily small probability ε. Paul Samuelson (1963) called this the "epsilon ain't zero" problem.

A second problem arises: if risk aversion (as measured by the concavity of the utility function in wealth) is used to explain why people turn down bets over stakes in the 0 to $1,000 range, then it cannot possibly explain why virtually *any* bets are accepted over large stakes. An economist who had observed an individual reject the opportunity to flip a coin to either win $1010 or lose $1000 would invoke risk aversion as the explanation. But Matthew Rabin (2001) pointed out that the level of risk aversion necessary to explain this choice would also imply that the same individual would turn down a coin flip for either an $80,000 loss or a $349,400 gain. The problem is that for small stakes, a concave utility function is approximately linear, and the amount of concavity necessary to explain why small stakes bets are sometimes rejected implies that most bets over large stakes—even very lucrative ones in expected value terms—would never be accepted.

The idea that sharply diminishing marginal utility of wealth arising from a concave utility function would disincline an individual from risk taking over large stakes is surely correct. But the two problems above suggests that concavity alone cannot explain behavior in the face of risk. The first is familiar: the conventional approach abstracts from loss aversion. The second is deeper: even if the utility function were continuously differentiable (not kinked at the status quo state, as would be the case if loss aversion were present), its concavity fails to capture the reasons people have for wishing to avoid risk and the emotions they experience in the face of risk. Among these are anxiety and fear when they do not know what will happen or the possibility of regret (or shame) at having taken a chance which ex post did not pay off. The model correspondingly fails to understand the reasons why people of very limited wealth engage in risky activities such as gambling: it is unlikely that their utility functions are *convex* in wealth, and if they are, it then begs the question of why the same individuals also purchase insurance. A more plausible explanation of gambling, and of driving too fast, too, is that some people enjoy taking particular *kinds* of risks.

Situation-dependent utilities, as well as the specific shortcomings of the expected utility maximization approach to risk and the discounted utility approach to intertemporal choice, suggest that a more empirically grounded view of the reasons for behavior is called for. Daniel Kahneman, Amos Tversky, Richard Thaler, and their coauthors have suggested a series of reformulations called *prospect theory* (the key papers are presented in Kahneman and Tversky 2000). Its main contribution is to take account of four aspects of choice not well handled in the conventional paradigm. The first is the problem (mentioned above) that people do not evaluate risky decisions according to the expected utility hypothesis: they overweigh the importance of unlikely events. The second is to take account of *framing*, namely, the fact that equivalent outcomes are treated differently depending on the manner in which either the outcomes or the decision setting are described. One of the reasons for situation-dependent behavior is that situations often frame choices in a particular manner. (Examples will be given in the next section.) Third, Kahneman and others, returning to an aspect of classic utilitarianism, have reintroduced substantive measures such as actually experienced hedonic utility.

Fourth, prospect theory has developed a conceptual framework for dealing with the situation-dependence of behaviors. This fundamental reformulation is that if the utility function is to explain actual behavior, its arguments should be *changes in states* or *events* rather than states. Thus, the value individuals place on states depends on the relationship of the state to the status quo (or possibly some other reference state,

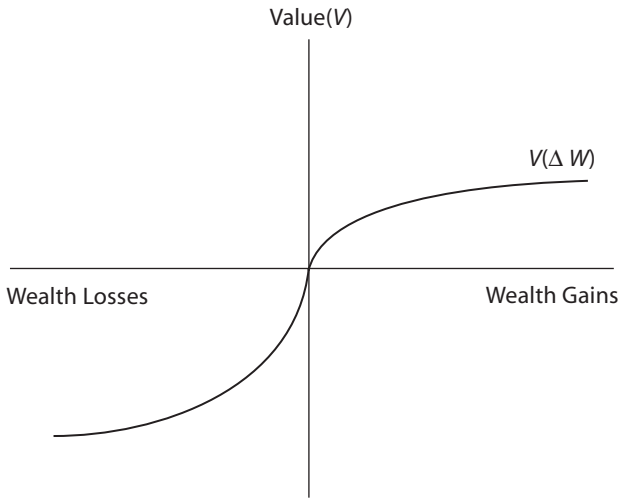

Figure 3.1 A situation-dependent value function. ΔW is the change in wealth. Note: the 'kink' at $\Delta W = 0$ indicates loss aversion.

such as an aspiration level or the states enjoyed by peers). Experimental and other empirical studies suggest that the resulting so-called value function has the three characteristics illustrated in figure 3.1, namely, that value is defined on changes in wealth rather than levels, that the value function is "kinked" at the status quo with a loss aversion coefficient of about two or a bit more (the function immediately to the left of the status quo is twice as steep as to the right), and that the marginal value of changes is diminishing with larger deviations from the status quo having ever smaller marginal effects on the individual's valuation of the event so that the value function is convex in losses and concave in gains (called the *reflection effect*).

A shortcoming of this literature to date, as the loss aversion interpretation of the stock premium puzzle revealed, is that situation dependence is an incomplete representation of preferences unless it is joined with an account of how preferences dynamically adjust to new situations, which is to say, how preferences evolve. The experiments and other data introduced here show that situations induce preferences; but they tell us little about the process by which people adjust to a new situation, whether it be stock market losses, the lost of sight, the promotion into a managerial position, or the transformation of a simple hunter-gatherer society to a modern market-based economy. I will return to the evolution of preferences under the influence of changing economic situations in chapter 11.

The shortcomings and reformulation considered in this section have addressed the formal core of the conventional theory of rational action. The recent accumulation of empirical anomalies concerning the substantive aspect of the theory, namely, the axiom of self-interested behavior, has also motivated reformulations based on the concept of social preferences.

Social Preferences

In one-shot prisoners' dilemma experiments, the rate of cooperation is commonly between 40 and 60 percent, despite mutual defection being the dominant strategy equilibrium (Fehr and Fischbacher 2001b). Many subjects prefer the mutual cooperation outcome over the higher material payoff they would get by defecting on a cooperator. When they defect, it is because they hate being taken advantage of; many defect to avoid risking this, not because it is the payoff maximizing strategy independently of the other's actions. These results suggest that people care about others, and they care about why things happen independently of the outcome. *Social preferences* are these *other-regarding* and *process-regarding* reasons for behavior.

Here is an example of *process-regarding preference*: you may accept with equanimity a bad outcome determined by a coin flip, while angrily refusing the outcome were it imposed by someone whose intention was to harm you. A process-regarding preference is defined as an evaluation based on the reasons why a state occurred rather than any intrinsic characteristic of the state. Other examples include a desire to help the less well off only if their poverty is the result of bad luck rather than laziness, keeping promises, and a predisposition to share things acquired by chance but not those acquired by one's effort. *The key aspect of process-regarding preferences is that the evaluation of a state is conditional on how it came about*. Behaviors are process sensitive for two reasons: the processes that determine an outcome often reveal important information about the intentions of others (e.g. the deserving poor), and they often provide cues concerning socially appropriate behaviors.

Other-regarding preferences include spite, altruism, and caring about the relationship among the outcomes for oneself and others. What Hobbes called the desire for "eminence" or a preference for "fair" outcomes are examples, as is Thorsten Veblen's "pecuniary emulation" exemplified by a desire to "keep up with the Joneses" (Veblen 1934 [1899]). *The key aspect of other-regarding preferences is that one's evaluation of a state depends on how it is experienced by others*. In analyzing preferences defined over the experiences of others (as well as one-

TABLE 3.1
A taxonomy of behaviors: costs and benefits to self and others

	Cost to self	*Benefit to Self*
Benefit to other	Altruism	Mutualism
Cost to other	Spite	Selfish

self), it will be helpful to consider the following taxonomy (see table 3.1) of the distribution of benefits and costs when two people interact.

The left-hand column lists behaviors that are specifically precluded by the self-interest axiom. A behavior is *altruistic* if it confers a benefit on another while inflicting a cost on oneself (this standard biological definition is restricted to benefits and costs and does not concern intentions). Inflicting a cost on another at a cost to oneself (the lower left) may be motivated by spite, envy, inequality aversion (if the other is richer), or the desire to punish those who have done harm to you or to others or who have violated social norms. The right-hand column is familiar territory for economists. Because in the conventional model market exchange is undertaken for self-interested reasons, it must confer benefits on both parties and hence is an example of what biologists call *mutualism* (when it occurs between members of different species). Other examples include seemingly generous behaviors that increase an individual's payoffs over the long term due to repeated or indirect interactions. Following Robert Trivers (1971) these behaviors are sometimes called "reciprocal altruism," a misnomer given that the reciprocal altruist benefits from the behaviors in question. The Dalai Lama's terminology is more accurate: "The stupid way to be selfish is . . . seeking happiness for ourselves alone. . . . The intelligent way to be selfish is to work for the welfare of others" (Dalai Lama 1994:154). I restrict the term self-interested to the behaviors in the right column to avoid the tautological use of the term to mean any act that is voluntarily undertaken. The altruist may give with pleasure, but clarity is not served by calling this self-interest.

Everyday observation of others as well as introspection suggests that other-regarding and process-regarding preferences are important. I will shortly introduce experimental evidence that confirms these impressions. But I want to stress that the main evidence for social preferences comes not from experiments but from real world economic and other behaviors that are inexplicable in terms of self-interest (without resort to extensive ad hoc reasoning). Some of these behaviors were referred to in the introduction of this chapter. Others include volunteering for dan-

gerous military and other tasks, tax compliance far in excess of that which would maximize expected incomes (in some countries), participating in various forms of collective action, and conforming to norms and laws in cases in which one's transgression would not be detected. Humans are unique among animals in the degree to which we cooperate among large numbers of non-kin; some of this cooperation is surely the result of institutions that make cooperative behavior a best response for people with self-regarding preferences (making cooperation a form of mutualism), but nobody seriously thinks that *all* of it can be explained this way.

There is an extensive literature on altruism, social comparison and other aspects of social preferences. I will illustrate the importance of social preferences by reference to *strong reciprocity*, not to be confused with the self-interested behaviors described by Trivers's "reciprocal altruism" and related concepts such as "indirect reciprocity" (conferring benefits on those who have benefitted others and receiving benefits in return as a result). By contrast to these "intelligent ways of being selfish," strong reciprocity motives may induce behaviors that are altruistic in the biologists' sense, conferring benefits to others in one's group at a cost to oneself. But reciprocity differs from altruistic behavior, which is not conditioned on the type or actions of the other.

The commonly observed rejection of substantial positive offers in the experimental Ultimatum Games is an example of reciprocity motives. Experimental protocols differ, but the general structure of the Ultimatum Game is simple. Subjects are anonymously paired for a single interaction. One is the "responder," and the other the "proposer." The proposer is provisionally awarded an amount ("the pie," "the pot," or some other culinary metaphor) known to the responder to be divided between proposer and responder. The proposer offers a certain portion of the pie to the responder. If the responder accepts, the responder gets the proposed portion and the proposer keeps the rest. If the responder rejects the offer, both get nothing. Figure 3.2 presents a version of the game in extensive form, with A's payoffs first. In this version the proposer simply chooses between two offers: divide the pie equally (5,5) or keep 8 and offer the respondent 2.

In this situation, the self-interest axiom predicts that an individual's actions are best responses defined over the outcomes of the game based on beliefs that other players also conform to the self-interest axiom. The self-interested proposer A will (by backward induction) determine that the responder B will accept the offer of 2 (because A believes that B is also self-interested) and so will propose the 8,2 split, which B will accept. In games in which an offer lower than 2 is possible, the self-interest axiom predicts that the proposer will offer either zero or the smallest

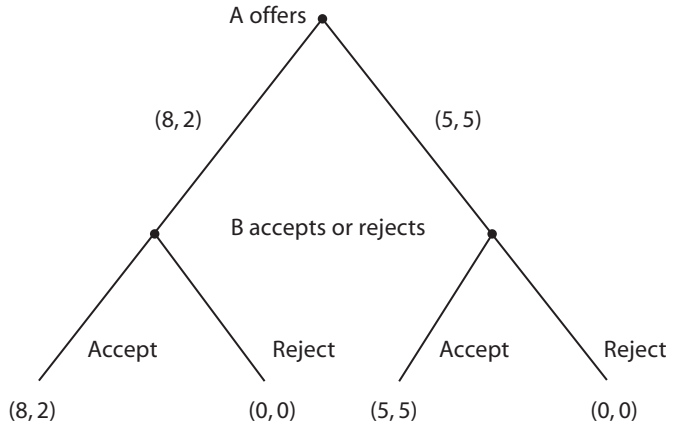

Figure 3.2 An ultimatum game. Note: Unlike the typical game, A's offer is restricted to just (5, 5) or (8, 2).

possible amount (in most games, the proposer can choose all values in whatever unit the pie is denominated from zero to the entire pie).

This game has been played anonymously for real money in hundreds of experiments with university student subjects in all parts of the world. The prediction of the self-interest axiom invariably fails. Modal offers are typically half of the pie, mean offers generally exceed 40 percent of the pie, and offers of a quarter or less are rejected with probabilities ranging from 40 to 60 percent. In experiments conducted in the United States, Slovakia, Japan, Israel, Slovenia, Germany, Russia, Indonesia, and many other countries, the vast majority of proposers offer between 40 and 50 percent of the pie (Fehr and Gaechter 2000b).

These results are interpreted by many as evidence for reciprocity motives on the part of the responder, who is willing to pay a price (forgoing a positive payoff) to punish the proposer for making an offer deemed unfair. The behavior of proposers is more complicated. Whether the large number of even splits (and other seemingly fair or near-fair offers) is explained by adherence to fairness norms or altruism by the proposer or to self-interestedness informed by a belief that the responder may reject an unfair offer cannot be easily determined. Substantial offers violate the self-interest axiom in *either* case, but the proposer does not exhibit reciprocity for the simple reason that as first mover he has no information about B on which to condition his behaviors. The evidence for reciprocity motives thus comes from the responders' behaviors, not the proposers'. Other interpretations—the respondents may be seeking to implement an egalitarian outcome rather than to punish the proposer, for example—have been suggested, but as we

will see presently, the evidence for reciprocity motives is quite compelling.

Results challenging the fundamental behavioral model in economics were bound to be subjected to critical scrutiny. Some wondered if the results were due to the relatively low stakes in the game. But subsequent experiments conducted among university students in Indonesia for a "pie" equal to three months average expenditures reproduced the same results (Cameron 1998). Experiments with U.S. students with a pie of $100 to $400 did not alter the results (Hoffman, McCabe, and Smith 1996, Fehr and Fischbacher 2001b). Behavior consistent with social preferences has been common in other high-stakes game—for example, a gift exchange game in Russia with earnings two- to three-times the monthly income of the subject (Fehr and Fischbacher 2001b). It appears that the violations of the predictions of the standard model are not the result of the stakes being too small to focus the attention or elicit the true motives of the experimental subjects. Others suggested that subjects' may have misunderstood the game, but later experiments in which subjects played the game many times with different partners lent no credence to this concern (Fehr and Fischbacher 2003). A final skeptical suggestion was that the subjects may not have adapted their behavior to the nonrepeated nature of the interaction, perhaps following rules of thumb derived from more common repeated interactions. But experimental subjects readily distinguish between repeated and nonrepeated interactions (adapting their behaviors accordingly). And in any case, use of a rule of thumb consistent with the observed play contradicts the standard model, however it came about. While debate concerning the interpretation of the games continues, there is a consensus that other-regarding motives are involved.

That other-regarding motives are important is not the only lesson. Suppose the ultimatum game in figure 3.2 were to be played with slight modifications in the protocol. In the experiment called Γ_1, the designation of proposer (occupied by *A* in the figure) is determined, as it is in most experiments, by a coin flip: if the coin says that *A* is the proposer, the game is as in figure 3.2. In Γ_2 the proposer is selected as in Γ_1 but a second coin is then flipped determining which action *A* will take. *A* then makes the indicated offer and finally *B* rejects or accepts. Introspection, as well as experimental results, suggest that the two games are fundamentally different in the behaviors they will evoke from *B*, even though *B* is choosing among the identical payoffs in both. In games like Γ_2, low offers are accepted that in Γ_1 would be rejected. A plausible explanation of the difference concerns reciprocity. In Γ_2 reciprocity motives will not come into play because *B* knows that, should the coin flip dictate the 8,2 proposal, *A* did not *intend* to make an unfair offer but was merely

TABLE 3.2
Varieties of ultimatum game play

Game	*Results*	*Interpretation*	*Source*
Γ_1 Standard	Modal offer ½, offers < 20% rejected	Reciprocity by respondent	Cited in text
Γ_2 Randomized offers	Few rejections of low offers	Proposer not responsible	Blount (1995)
Γ_3 Roles chosen by quiz	Many low offers, few rejections	Proposer "deserving"	Hoffman, McCabe, Shachat, and Smith (1994)
Γ_4 "Exchange Game"	Many low offers, few rejections	Situational framing	Hoffman, McCabe, Shachat, and Smith (1994)
Γ_5 No "fair" offers possible	Low offers not rejected	Proposer's intentions matter.	Falk, Fehr, and Fischbacher (2003)
Γ_6 Punishment by third party	C punishes A's low offer to B	Generalized fairness norms	Fehr and Fischbacher (2001a)
Γ_7 Standard: Au/Gnau	Offers > ½ common and are rejected	Endogenous and situation-dependent prefs	Henrich, Bowles, Boyd, Camerer, Fehr, Gintis, and McElreath (2001)
Γ_8 Standard: Machiguenga	Many low offers, very few rejections	Endogenous and situation-dependent prefs	Henrich (2000)

constrained to do so by the rules of the game. The comparison illustrates process-regarding preferences: in both cases *B* got a bad offer, but in the second case the process determining the bad deal made it clear that it did not come about as a result of *A*'s bad intentions. Had rejections of low offers in Γ_1 been motived by inequality aversion rather than reciprocity motives, for example, Γ_2, would have been played the same way.

Now consider Γ_3, in which the proposer position is designated not by a random draw but on the basis of a current events quiz taken prior to the play of the game, with *A*, the higher scorer, becoming the proposer, to whom *B* responds. Experimental subjects play Γ_3 differently than the

standard Γ_1: proposers are more likely to keep a substantial fraction of the pie to themselves, and quite unequal proposals are frequently accepted. Now alter the game again, this time simply by naming Γ_4, "The Exchange Game" rather than "Divide $10." What the game is *called* should have not effect on behavior in the conventional framework, but it does: proposers offered less and lower offers were accepted. These and other experiments are summarized in table 3.2.

It is not difficult to think of reasons why people play Γ_3 differently from Γ_1: responders may feel that the proposers' low offers should not be punished as they reflect the proposers' greater deservingness (having earned the role of proposer by their test score). But what of Γ_4, "The Exchange Game"? It seems likely that the experimental subjects took the name of the game as a situational cue and as a result acted on the more self-regarding motivations in their behavioral repertoires. But however one understands the differences, they cannot be accounted for by the structure of the game's payoffs, for these remain unchanged by the differing processes of role designation, framing, and selection of actions. Another variant of the game (Γ_5) reaffirms the impressions that rejections are motivated by the desire to punish unfairness on the part of the proposer, not simply by the desire to avoid accepting an uneven split: retain the 8,2 offer of the standard game, but restrict the proposer to 10, 0 (an even more "unfair" offer) as the only alternative to 8,2. Rejections of the 8,2 offer were less than a quarter as frequent in Γ_5 than in Γ_1.

An important role for ethical values is suggested by Γ_6, which involves three people and is not strictly an Ultimatum Game. *A* assigns some part of the pie to *B* (who simply receives the offer and has no other role); then *C*, who has observed the pie size and the offer, may choose to reduce *A*'s payoff by allocating some of *C*'s endowment (like the pie provided by the experimenter) for this purpose. Allocations by A of half or more of the pie to *B* are never punished; but when *A* gives *B* less than half, *C* is willing to pay to punish *A*. In this case *C* acts very much like the responder in the standard Ultimatum Game but is responding to a seemingly unfair offer not to himself but to another (anonymous) person. Fehr and Fischbacher found than punishment by such third parties as *C* is only slightly less strong than punishment by the recipient of a low offer in the standard ultimatum game setup.

I report also two experiments in which the subject pool is not—as is usually the case—composed of university students but instead were members of fifteen small-scale societies with little contact with markets, governments, or modern institutions. My colleagues and I (a team of 17 anthropologists and economists) designed the experiments to explore whether the results reported above are common in societies with quite

different cultures and social institutions (Henrich, Bowles, Boyd, Camerer, Fehr, Gintis, and McElreath 2004). The fifteen societies included hunter-gathers, herders, and farmers. Among the Au and Gnau people in Papua New Guinea, offers of more than half of the pie were common, and high and low offers were rejected with equal frequency. This seemingly odd result is not surprising in light of the practice of competitive gift giving as a means of establishing status and subordinacy in these and many other New Guinea societies. By contrast, among the Machiguenga in Amazonian Peru, almost three-quarters of the offers were a quarter of the pie or less and there was just a single rejection, a pattern strikingly different from the experiments conducted thus far. However, even among the Machiguenga, the mean offer was 27 percent, suggesting that offers exceeded the expected payoff maximizing offer.

Analysis of the experiments in the fifteen simple societies we studied led us to the following conclusions: behaviors are highly variable across groups, not a single group approximated the behaviors implied by the self-interest axiom, and between group differences in behavior seem to reflect differences in the kinds of social interaction experienced in everyday life. The evidence for economic conditions affecting behavioral norms is quite compelling. For example, the Aché in Paraguay share equally among all group members some kinds of food (meat and honey) acquired through hunting and gathering. Most Aché proposers contributed half of the pie or more. Similarly, among the Lamalera whale hunters of Indonesia, who hunt in large crews and divide their catch according to strict sharing rules, the average proposal was 58 percent of the pie. Moreover the Indonesian whale hunters played the game very differently from the Indonesian university students mentioned above.

The Ultimatum Game is one of many in which experimental subjects have behaved in ways that are strongly at variance with the predictions of the self-interest axiom. Colin Camerer and Ernst Fehr (2004) survey seven games in which experiments have suggested the salience of social preferences. One of these, the Public Goods Game, is both important as an analogy to many real world economic problems, and instructive about human behavior. It is sometimes called an *n*-person prisoners' dilemma because it has the same incentive structure: if players conform to the self-interest axiom, contributing nothing to the public good (analogous to defection) is the dominant strategy equilibrium, but universal contribution maximizes total payoffs. Here is the game: n players are each given an "endowment" y and then simultaneously select an amount $c_i \in [0,y]$ to contribute to the public good. Each player's payoffs are $\pi_i = y - c_i + m\Sigma_j c_j$ for $j = 1 \ldots n$. This describes a Public Goods Game if $m < 1 < mn$. The first of these inequalities implies that the individual's best response is to contribute nothing, and the second

implies that total payoffs (summing over the group) are maximized if everyone contributes his entire endowment. Variants of the Public Goods Game have been used to model individual extraction from a common pool environmental resource; applications include contribution to joint projects such as payment of taxes and participating in strikes.

The prediction of the self-interest axiom ($c_i = 0 \; \forall \; i$) is consistently contradicted in experiments (surveyed by Ledyard 1995). In one-shot games contributions average about half of the endowment, while in multi-period games contributions begin around half and then decline, so that a majority of players contribute nothing in the final round of a ten-round game. This decline in contribution was initially thought to confirm the conventional model, the idea being that once the subjects had figured out the game, they contributed nothing. But an experiment in which a second ten-round public goods game was unexpectedly begun at the end of the first ten-round game suggests that this is not the case: in the second game players again began with contributions of about half. Many have interpreted the decline in contributions as a reflection of the disappointment of expectations that others would contribute more, along with the desire to punish low contributors (or at least not to be taken advantage of) in a situation in which this could only be done by reducing one's own contributions.

Strong support for this latter view is provided by an ingenious experiment designed by Fehr and Gaechter (2002): it has the same public goods structure as above except that after individuals contributed, their contributions were made known (by an identifying number only, not by name, of course) to all group members, who then had the opportunity to punish others in the group, imposing a cost (payoff reduction) on both the punisher and the punished.[9] In one experiment with this game, Fehr and Gachter adopted what is called the perfect strangers treatment: after each round of the ten-round experiment the groups were reshuffled so that players knew that no player would ever encounter any other player more than once. Thus, the motivation for punishment could not be self-interest. If low contributors respond to punishment by contributing more in subsequent rounds, they raise the payoffs of others but not of the punisher (due to the perfect strangers treatment). Thus punishment is no less a public good than the initial contribution. This is transparently the case on the last round of the game, when the last action taken by any player is the decision to engage in costly punishment of fellow group members: those who punish under these condi-

[9] An earlier experiment of this type with similar results is Ostrom, Gardner, and Walker (1994).

tions must value punishment per se rather than any anticipated consequence of punishment for their game payoffs based on the modification of the behaviors of others.

In Fehr and Gaechter's Public Goods Game with punishment, contributions started at about half the endowment (as in the standard game) but then *rose* rather than fell over the course of the game. My coauthors and I (reported in Bowles and Gintis 2002b) implemented a similar game in which we confirmed what one would expect: punishment is directed at low contributors, and they respond strongly to punishment. Those who thought they could cheat on the last round by reducing their contributions paid dearly for their mistake. We also found something quite unexpected. When those contributing *above* the mean were punished (as they occasionally were), they sharply *reduced* their contributions. Even more striking is the fact that the positive response to punishment by the low contributors was not a best response defined over the game payoffs. Taking account of the observed relationship between the expected amount of punishment and one's offer, zero contribution remained the best response, but nonetheless those punished responded by contributing more.

A reasonable interpretation of these experiments is that, as in the Ultimatum Game, people are willing to pay to punish those who violate social norms even when there is no expectation of future or indirect payoff. In other words the subjects were acting in accordance with reciprocity motives. But something else seems to be at work. The fact that punishment induced more contribution by the shirkers (contrary to the payoff-maximizing choice, even when the likely punishment is taken into account) suggests that social sanction by peers may mobilize feelings of shame in situations in which the punishment carries some legitimacy (in the eyes of the person punished). In two similar experiments—one in the laboratory and one in the field among farmers in Zimbabwe—"punishment" merely conveyed displeasure and did not reduce the payoffs of the one punished. But the fact that those punished contributed more in subsequent periods shows the strong effects of social sanction, consistent with the "shame" interpretation (Barr 2001, Masclet, Noussair, Tucker, and Villeval 2003). In chapter 4 I provide a model of how social preferences such as shame and reciprocity may support cooperation in public goods interactions.

The Public Goods Game provides a nice example of situation-dependent behaviors and framing. Jean Ensminger conducted public goods experiments with the Orma, a herding people in Kenya, as part of the multi-cultural experimental project mentioned above. When the Orma need some public good—a new primary school or the repair of a road, for example—members of the community are asked for a voluntary

contribution to the project, the amounts increasing in the amount of wealth (cattle) of the family. This system of voluntary public goods provision is called *harambee*. When Ensminger explained the Public Goods Game to her subjects, they promptly dubbed it the "*Harambee* Game," and their contributions were strongly predicted by their (real world) wealth, just as would have been the case in a real *harambee*. When the Orma subjects played the Ultimatum Game, they did not analogize it to the *harambee* (or apparently to any other aspect of their everyday life) and wealth did not predict any aspect of their experimental play.

Do people behave in natural settings the way they do in experiments? The relationship between experimental play and real world behaviors is complex, and I do not want to claim an overly close correspondence between the two. Contrary to the (misguided, in my view) hopes of some experimenters, experimental games do not tap abstract motives uncontaminated by situations. In this, experimental play is much like any other behaviors and the experiment is just another situation.[10] The game situation, the instructions of the experimenter, and the like are a very strong frame and we cannot expect them to be without effect. Experiments do not reveal the essence of a universal human nature. Rather, they simply show that common behaviors in generic social interactions are readily explained by social preferences, thus suggesting that the many real world examples of seeming violations of the self-interest axiom are not the result of the peculiarities of the particular real world examples.

An Empirically Based Social Preference Function

In response to the violations of the self-interest axiom in a number of experiments, economists have attempted to reformulate a utility function capable of explaining the above behaviors in a parsimonious manner. Is there a utility function that is at once simple enough to be tractable and sufficiently robust to explain not just one of the experimental anomalies but all of them? There now exist a number of utility functions that are capable of explaining a wide range of experimental behaviors (Falk and Fischbacher 1998, Fehr and Schmidt 1999, Bolton and Ockenfels 1999, Rabin 1993, Charness and Rabin 1999, Levine

[10] Loewenstein (1999) provides a skeptical but balanced assessment. Behaviors in games have been shown to predict real world behaviors in a few cases: those who trusted in a trust experiment by Glaeser, Laibson, Scheinkman, and Soutter (2000), for example, exhibited more trust in a number of real world situations. By contrast, answers to standard survey questions on trust were completely uncorrelated with any measured behaviors (experimental or non-experimental).

1998). The basic ingredients of the proposed utility functions are self-interest, altruism, spite, fair-mindedness, and reciprocity. The functions differ in the way that these components are combined, and the types of behaviors the authors wish to stress.

Here is a utility function (proposed by Fehr and Schmidt) that takes account of both self-interest and what they term "inequality aversion." A fair (i.e., inequality averse) utility function of person i (interacting with just one other person, j) is given by

$$U_i = \pi_i - \delta_i \max(\pi_j - \pi_i, 0) - \alpha_i \max(\pi_i - \pi_j, 0) \qquad (3.3)$$

where π_j and π_i are the material payoffs to the two individuals, and $\delta_i \geq \alpha_i$ and $\alpha_i \in [0,1]$. This utility function expresses individual i's valuation of her own payoff as well as her aversion to differences in payoff, with disadvantageous differences ($\pi_j - \pi_i > 0$) being more heavily weighted (δ_i) than advantageous differences (α_i). The upper bound on α precludes what might be termed "self-punishing" levels of aversity to advantageous inequality: an individual with $\alpha = 1$ cares only about the other's payoffs (if they fall short of his own). By contrast, a person (i) very averse to disadvantageous inequality might prefer $\pi_j = \pi_i = 0$ to $\pi_i = 1$ and $\pi_j = 2$, so δ may exceed 1.

To see the implications of fair-mindedness for both sharing and punishing behaviors, suppose the two are to divide one unit ($\pi_i + \pi_j = 1$) and that $\alpha_i > \frac{1}{2}$. In this case $dU_i/d\pi_i < 0$ for all divisions such that $\pi_i - \pi_j > 0$. Thus individual i's preferred share would be to divide the unit equally (so if the share initially favored i over j, i would prefer to transfer some of the payoff to j). Similarly, if $\delta_i \geq \frac{1}{2}$ and payoffs were divided so that j was to receive 0.6 and i 0.4, i would be willing to pay 0.1 to reduce the payoffs of j by 0.3 so that both received 0.3. Even more striking, in this case, i would refuse an offer of less than 0.25 if by doing so both would receive nothing (as in the Ultimatum Game).

Fair-mindedness may explain another experimental anomaly mentioned at the outset: a substantial number of experimental subjects in one-shot prisoners' dilemma games cooperate (despite defecting being the dominant strategy in the game payoffs). A fairminded row player (one with the above Fehr-Schmidt utility function) facing the standard prisoners' dilemma material payoffs $a > b > c > d$ would cooperate if he knew the column player would cooperate as long as the disutility he experienced from advantageous inequality is sufficiently large, or $\alpha > (a - b)/(a - d)$ (see table 3.3).

If this inequality obtains (which it may because the right-hand side is necessarily less than unity), then the resulting game is no longer a prisoners' dilemma but rather an Assurance Game, so there exists some critical value $p^* \in (0,1)$ such that if Row believes that Column will

TABLE 3.3
Standard Prisoners' Dilemma and Fair-minded Utility Payoffs for Row

	Cooperate	*Defect*
Cooperate	b	d
	$\boldsymbol{b}$	$\boldsymbol{d - \delta(a - d)}$
Defect	a	c
	$\boldsymbol{a - \alpha(a - d)}$	$\boldsymbol{c}$

Note: utility payoffs for fairminded row player are in bold.

defect with probability less than p^*, then his best response is to cooperate. You can also readily show that $dp^*/d\alpha > 0$ while $dp^*/d\delta < 0$, so if this interaction took place among randomly paired fairminded players in an evolutionary setting of the type modeled in the previous chapter, increasing the disutility of advantageous inequality enlarges the basin of attraction of the mutual cooperate equilibrium while increasing the disutility of disadvantageous inequality does the opposite.

In an experiment designed to estimate the parameters of a function like eq. (3.3) Loewenstein, Thompson, and Bazerman (1989) created a variety of scenarios that had in common that an amount had to be divided, but the situations differed in the personal relationship among the participants (negative, neutral, or positive) and in the nature of the interaction (business, other). They found that disadvantageous inequality was strongly disliked, irrespective of the nature of either the personal relationship or the transaction. By contrast, advantageous inequality was disliked by 58 percent of the subjects in the nonbusiness transaction but was preferred by most in the business transaction, being disliked by only 27 percent. The nature of the personal relationship mattered, too: in the positive personal or neutral relationship setting, 53 percent disliked advantageous inequality, while in the negative relationship setting only 36 percent did. This experiment provides direct evidence on inequality aversion and is also consistent with the view that behaviors are commonly conditioned on one's belief about the other person (positive or negative) and are situationally specific (business or not).

Fairminded preferences are defined over outcomes, but reciprocal preferences depend as well on one's belief about the intention or type of the individual one is dealing with. Following ideas initially laid out by Rabin (1993) and Levine (1998), the following function incorporates self-interest, altruism, and reciprocity. An individual's utility depends on

his own material payoff and that of other individuals $j = 1 \ldots n$ according to

$$U_i = \pi_i + \Sigma_j \beta_{ij} \pi_j \qquad \text{for } i \neq j \tag{3.4}$$

where β_{ij}, the weight of j's material payoff in i's preferences, is

$$\beta_{ij} = \frac{a_i + \lambda_i a_j}{1 + \lambda_i} \; \forall \; j \neq i \tag{3.5}$$

and $a_i \in [-1,1]$ and $\lambda_i \geq 0$. The parameter a_i is i's level of unconditional good will or ill will (altruism or spite) toward others, and $a_j \in [-1,1]$ is i's belief about j's good will, while λ_i indicates the extent to which i conditions his evaluations of others' payoffs on (beliefs about) the other's type. If $a_i = 0$ and $\lambda_i > 0$, then individual i is a nonaltruistic reciprocator (exhibits neither good will nor spite unconditionally but conditions her behavior on the goodness or spitefulness of others).

If $\lambda_i = 0$ and $a_i \neq 0$, then i exhibits unconditional altruism or spite, depending on the sign of a_i. The denominator is augmented by λ_i so that $\beta_{ij} \leq 1$, thereby restricting one's valuation of the others' payoffs to being no greater than one's own. Note that $d\beta_{ij}/d\lambda_i$ has the sign of $(a_j - a_i)$, which means that the level of reciprocity affects the extent to which others' payoffs enters into one's own evaluation, increasing it if the other is kinder than oneself, and conversely. If $a_j = a_i$ then $\beta_{ij} = a_i$ for any level of reciprocity.

Like the inequality-averse function, this reciprocity-based utility function can be used to explain generous and punishing behaviors. The analysis is considerably more complicated, however. In most social interactions we have some prior beliefs about the others' types based on knowledge of their prior behavior, cues based on other facts about them (including their status as an "insider" or an "outsider" in the current interaction), and the situation itself. Thus one's beliefs about the others' types and hence one's valuation of their benefits plausibly depends on their past actions, which depend on their beliefs on one's own type, and so on. If one is a reciprocator and believes that others are altruistic, one may engage in conditional generosity. But if the generosity is not reciprocated, one may update one's beliefs about the others' types and engage in punishment or at least withdrawal of generosity, as was witnessed in the public goods experiments. Thus, behaviors may be both path dependent and situationally specific: a situation that induces beliefs that others are altruistic may support high and sustainable levels of generosity, while the same individuals interacting in another situation may engage in mutually costly spiteful punishment. The path-dependent and situationally specific nature of behaviors may explain why subjects' play is so affected by changes in experimental protocols that would be irrele-

vant were the conventional model correct. It also might illuminate why such large differences in behaviors are found in our cross-cultural study.

Conclusion

The inequality-averse and reciprocity-based functions just presented are important steps toward the construction of a more adequate conception of behavior. But the process is ongoing and far from completion. The evidence that inequality aversion and reciprocity motives are common does not suggest that people are irrational. Indeed, strong experimental evidence indicates that when individuals give to others (e.g., in a Dictator Game) their behavior conforms to the transitivity assumptions and other requirements of rational choice (Andreoni and Miller 2002). Moreover, people respond to the price of giving, giving more when it costs them less to benefit the other. The importance of other-regarding motives thus does not challenge the assumption of rationality but rather suggests that the arguments of the utility function should be expanded to account for individuals' concerns for others.

The experimental and other evidence also suggests an adequate formulation should take account of the behavioral heterogeneity of most human groups. Using data from a wide range of experiments, Ernst Fehr and Simon Gaechter estimate that between 40 and 66 percent of subjects exhibit reciprocal choices. The same studies suggest that between 20 and 30 percent of the subjects exhibit conventional self-regarding outcome-oriented preferences (Fehr and Gaechter 2000b, Camerer 2003). Loewenstein, Thompson, and Bazerman (1989) distinguished among the following types in their experiments:

> *Saints* consistently prefer equality, and they do not like to receive higher payoffs than the other party even when they are in a negative relationship with the opponent . . . *loyalists* do not like to receive higher payoffs in neutral or positive relationships, but seek advantageous inequality when in negative relationships . . . *Ruthless competitors* consistently prefer to come out ahead of the other party regardless of the type of relationships. (p. 433)

Of their subjects, 22 percent were saints, 39 percent were loyalists, and 29 percent were ruthless competitors (the rest could not be classified).

Thus, the objective of a reformulation of the behavioral foundations of economics should not be some new *Homo sociologicus* to replace *Homo economicus*, but a framework capable of taking account of heterogeneity. This task is essential because heterogeneity makes a difference in outcomes, but it is challenging because the effects are not adequately captured by a process of simple averaging. The outcome of

interaction among a population that is composed of equal numbers of saints and ruthless competitors will not generally be the average of the outcomes of two populations with just one type, because small differences in the distribution of types in a population can have large effects on how *everyone* behaves.

Moreover, seemingly small differences in institutions can make large differences in outcomes. Imagine a one-shot Prisoners' Dilemma Game played between a self-interested player (for whom Defect is the dominant strategy in the simultaneous moves game) and a reciprocator (who prefers to Cooperate if the other cooperates and to Defect otherwise) (Fehr and Fischbacher 2001b). Suppose the players' types are known to each. If the game is played simultaneously, the reciprocator, knowing that the other will Defect, will do the same. The outcome will be mutual defection. If the self-interested player moves first, however, she will know that the reciprocator will match whatever action she takes, narrowing the possible outcomes to {Cooperate, Cooperate} or {Defect, Defect}. The self-interested player will therefore cooperate and mutual cooperation will be sustained as the outcome. Recall, as another example, that in the Public Goods-With-Punishment Game, those with reciprocal preferences not only acted generously themselves, but they apparently also induced the selfish types to act *as if* they were generous. But had there been too few reciprocators, all players (reciprocators and self-interested types alike) would have converged to zero contribution.

In addition to heterogeneity across individuals, versatility of individuals must also be accounted for. In the Ultimatum Game, proposers often offer amounts that maximize their expected payoffs, given the observed relationship between offers and rejections: they behave self-interestedly *but expected responders not to*. Moreover, *the same individuals* when in the role of responder typically reject substantial offers if they appear to be unfair, thus confirming the expectations of the proposer and violating the self-interest axiom.

Finally, as we have noted earlier (and will discuss in chapter 11), preferences are to some extent learned rather than exogenously given: durable changes in an individual's reasons for behavior often take place as a result of one's experience. This means that populations that experience different structures of social interaction over prolonged periods are likely to exhibit differing behaviors, not simply because the constraints and incentives entailed by these institutions are different but also because the structure of social interaction affects the evolution of both behavioral repertoires, the ways in which situations cue behaviors, and the way outcomes are evaluated. (Because the functioning of institutions depends on the preferences of the individuals involved, it will also be the case that institutions are endogenous with respect to preferences; I

model the resulting process, called the *coevolution of preferences and institutions*, in chapters 11 through 13.)

Progress in the direction of a more adequate behavioral foundation for economics must take account of these three aspects of people: namely, their *heterogeneity, versatility*, and *plasticity*.

New theories must also address two challenges. The first concerns the normative status of preferences. If preferences are to explain behaviors, they cannot unassisted also do the work of evaluating outcomes. The reason is that some common reasons for behavior—weakness of will, spite, and addiction come to mind—often induce behaviors the outcomes of which few would condone.

The second challenge arises because the experimental and other evidence indicating the importance of social preferences poses a difficult evolutionary puzzle. If many of us are fairminded and reciprocal, then we must have acquired these preferences somehow, and it would be a good check on the plausibility of social preference theories and the empirical evidence on which they are based to see if a reasonable account of the evolutionary success of these preferences can be provided. Generosity toward one's genetic relatives is readily explained. The evolutionary puzzle concerns nonselfish behaviors toward non-kin (meaning behaviors bearing individual costs with no benefit, or the lefthand column in table 3.1, above.) Among non-kin, selfish preferences would seem to be favored by any payoff-monotonic evolutionary processes, whether genetic or cultural. Thus, the fairmindedness that induces people to transfer resources to the less well off, and the reciprocity motives that impel us to incur the costs of punishing those who violate group norms, on this account, are doomed to extinction by long term evolutionary processes. If social preferences are common, this conventional evolutionary account must be incorrect.

In later chapters I return to this question and provide a series of models explaining the evolutionary success of social preferences. In particular I will explore the contribution to the evolutionary success of nonselfish traits made by characteristic structures of human social interaction, namely, social segmentation, repeated interactions, and reputation building (in chapter 7) and the enforcement of group-level norms and intergroup conflict (in chapters 11 and 13). In many cases the evolutionary success of what appear to be unselfish traits is explained by the fact that when an accounting of long-term and indirect effects is done, the behaviors are payoff-maximizing, often representing forms of mutualism. But I will also introduce plausible models accounting for the evolutionary success of behaviors that benefit other members of ones's group at a cost to oneself.

Like the theory of social preferences, prospect theory also raises evo-

lutionary puzzles. Hyperbolic discounters act in time-inconsistent ways; their average payoffs over a long period would be increased if they conformed to the dictates of the discounted utility model. Similarly, those who overweigh low probability events will earn lower expected payoffs than competitors who do the proper expected utility maximization. This does not mean that those using time-inconsistent discounting and violating the expected utility axioms are doomed, but given that either genetic or cultural evolution tends to favor those with higher payoffs, it does pose a puzzle. Similarly, loss-averse individuals forgo opportunities for substantial expected gains in risky situations. Their loss aversion thus disadvantages them in competition with others whose utility function is not kinked at the status quo. These evolutionary conundrums raised by prospect theory have received less attention than the puzzle of social preferences. I will not address them further, except to note that the initial evidence for hyperbolic discounting came from pigeons and rats, so this is not a uniquely human behavior.[11]

In chapter 4 I generalize the kinds of coordination problems introduced in chapter 1 as 2×2 games, and analyze the impressive variety of institutions, norms, and other ways people have developed to avoid or attenuate coordination failures. Social preferences, we will see, play a central role in this process.

[11] Hyperbolic discounting in humans and other animals is described in Ainslie (1975), Green and Myerson (1996), and Richards, Mitchell, de Wit, and Seiden (1997).