11

Homo reciprocans, *Homo egualis*, and Other Contributors to the Human Behavioral Repertoire

> The Americans…are fond of explaining almost all the actions of their lives by the principle of self-interest rightly understood… . In this respect I think they frequently fail to do themselves justice.
>
> *Alexis de Tocqueville*

> My motive for doing what I am going to do is simply personal revenge. I do not expect to accomplish anything by it… . Of course, I would like to get revenge on the whole scientific and bureaucratic establishment…but that being impossible, I have to content myself with just a little revenge.
>
> *Theodore Kaczynski (the Unabomber)*

11.1 Introduction

This chapter is probably the most important in the book in terms of the contribution of game theory to our understanding of what is specifically human about human sociality. It also represents an area of quite active contemporary research. While this material will be attractive to students, it presents a problem for instructors who, in all likelihood, have neither studied nor taught in this area. For this reason, I have presented the novel material at a more leisurely pace than the material in other chapters, and have supplied more descriptive details and a greater number of references to the literature. The first time one teaches this material, class presentations and discussions may be preferable to a formal lecture format.

This chapter has three main points. First, in many decision-making and strategic settings people do not behave like the self-interested "rational" actor depicted in neoclassical economics and classical game theory. Second, despite its increased complexity in comparison with traditional *Homo economicus*, human behavior can be modeled using game theory and optimization subject to constraints. Third, there are plausible models of human cultural and genetic evolution that explain how we have gotten to be the way we are. Our analytical and evolutionary models, however, leave considerable room for improvement, and we are presently on the steep portion of the learning curve in developing analytical models of human behavior.

We begin with an overview of the experimental method and its results, including methodological discussions concerning the interpretation of experimental results (§11.2.1) and the meaning of "rationality" (§11.2.2). While it is easy to discuss such topics in excess, I have found some class time devoted to these issues to be amply rewarded.

Section §11.3, Behavioral Economics: Games against Nature and against Ourselves, presents the results of some thirty years of brilliant research by Daniel Kahneman, Amos Tversky, and their associates into individual decision-making processes. This material is important for the microeconomics of individual choice but does not appear in our models of strategic interaction, so it can be relegated to assigned reading if the instructor is pressed for time. Section §11.4, subtitled The Laboratory Meets Strategic Interaction, is an overview of the basic empirical studies of social interaction upon which our analytical models are based and is worthy of careful study and discussion.

The next two sections represent attempts to model some of the experimental results. In *Homo egualis* (§11.5) we show that if individuals have inequality aversion, we can explain some experimental results, including why altruism appears in ultimatum and public goods games but not in marketlike interactions. In *Homo reciprocans:* Modeling Strong Reciprocity (§11.6) we model some of the experimental results that depend on the tendency of people to cooperate and punish as forms of prosocial behavior.

Economists are fond of explaining a complex social division of labor involving cooperation among individuals with low biological relatedness using models of complete contracting (neoclassical economics) or repeated games (classical game theory). Of course *reciprocal altruism* (Trivers 1971), which is self-interested cooperation using trigger strategies and reputation-building in a repeated game (§6.10, and chapter 6 in general) is important in humans, probably occurs a bit among primates (Byrne and Whiten 1988), and perhaps in a few other species (Pusey and Packer 1997). But as we suggest below, it is not plausible to model human sociality based on self-interested behavior alone.

This leads us to evolutionary models of non-*Homo economicus* behavior. Economists often argue that only self-interested behavior is evolutionarily viable. Yet living organisms routinely sacrifice resources that could be used for self-preservation for the sake of producing, nurturing, and protecting offspring(Daly and Wilson 1983). Self-interested agents are therefore rare mutants with low fitness and no evolutionary future. Perhaps then economists take "self-interest" to mean "family interest." Using William Hamilton's principle of *kin selection* we can indeed explain much of the apparent altruism in most species (Wilson 1975, Grafen 1984, Krebs and Davies 1993). But not so in humans.

In Altruism and Assortative Interactions (§11.7) we sketch an approach to the study of altruism that has proven illuminating and draws on standard notions from game theory and population biology. This approach is based on *Price's equation*, which, while intimately related to the replicator equation, is not yet widely known to economists. In The Evolution of Strong Reciprocity (§11.8) we apply Price's equation to a game-theoretic model explaining the evolutionary emergence of *Homo reciprocans*.

In *Homo parochius:* Modeling Insider/Outsider Relations (§11.9) we close with a model of the evolutionary emergence of ethnic and other preferences in which agents act in a distinctive prosocial manner with respect to "insiders" and correspondingly antisocial behavior toward "outsiders."

I have included few exercises in this chapter, in the belief that the material is new and exploratory and the student would do better to delve into some of the source material rather than solve problems. The instructor might assign additional readings from the references and ask students to write short papers based on such readings.

11.2 Modeling the Human Actor

When the same object of knowledge is modeled in different ways, the models should agree where they overlap—as, for example, in the smooth transitions from physics to chemistry to biology. This principle has been routinely violated in the social sciences, which maintain mutually incompatible theories across various disciplines. Economists hold that individuals are self-interested and maximize utility subject to constraints. Sociologists hold that individuals conform to societal norms. Social psychologists hold that individuals identify with groups and further their interests. Animal behaviorists derive behavior from the morphology and evolutionary history of a species. Correspondingly incongruous models are affirmed by anthropologists, psychologists, and political scientists.

This is a great scandal, and game theory is an important tool for moving beyond it.[1]

Game theory is a general lexicon that applies to the behavior of life forms, providing the tools for carefully modeling the conditions of social interaction, the characteristics of players, the rules of the game, the informational structure, and the payoffs associated with particular strategic interactions. This fosters a unified behavioral theory and also allows *experimental game theory* to use the same language and techniques, whether in biology, anthropology, social psychology, or economics. Since game-theoretic predictions can be systematically tested, the results can be replicated by different laboratories (Plott 1979, V. Smith 1982, Sally, 1995).

While many of the predictions of neoclassical economic theory (§3.17) have been verified experimentally, many others have been decisively disconfirmed. What distinguishes success from failure?

- When modeling market processes with well specified contracts, such as double continuous auctions (supply and demand) and oligopoly, game-theoretic predictions are verified under a wide variety of social settings (Davis and Holt 1993, Kachelmaier and Shehata 1992).[2]
- Where contracts are incomplete and agents can engage in strategic interaction, with the power to reward and punish the behavior of other players, the neoclassical predictions generally fail (§11.4).

In other words, *precisely* where standard neoclassical models do well *without* the intellectual baggage of game theory, game theory predicts well, but where game theory has something really *new* to offer, its predictions fail.

The culprit is the representation of the human actor—the so-called *rational actor model*—adopted by game theory. We will call this the *Homo economicus* model, because there is nothing particularly "rational" (or "irrational") about it. Neoclassical economics has accepted this model because when faced with market conditions—anonymous, nonstrategic interactions—people behave like self-interested, outcome-oriented actors

[1] Edward O. Wilson has mounted a powerful contemporary plea for the unity of the sciences. Wilson maintains that physics is the basis for such unity. However, since physics has no concept of strategic interaction, a concept central to all life forms, game theory is a more plausible unifying force for the behavioral sciences.

[2] Even here experimental economics sheds much new light, particularly in dealing with price dynamics and their relationship to buyer and seller expectations (Smith and Williams 1992).

(although probably not time consistent). In other settings, especially in the area of strategic interactions, people behave quite differently.

11.2.1 Interpreting the Results of Experimental Game Theory

When the results of experiments contradict received wisdom in economics, many economists reject the experiments rather than the received wisdom. For instance, in the ultimatum game (§11.4.1), individuals offered a small share of the pie frequently choose a zero payoff when a positive payoff is available. Critics claim that subjects have not learned how to play the game and are confused by the unreality of the experimental conditions, so their behavior does not reflect real life. Moreover, whatever experimentalists do to improve laboratory protocols (e.g., remove cues and decontextualize situations), the critics deem as insufficient, and the experimentalists complain among themselves that the critics are simply dogmatic enemies of the "scientific method."

To move beyond this impasse we must recognize that the critics are correct in sensing some fundamental difference between experiments in social interaction and the traditional experimental method in natural science, and that experimental results must be interpreted more subtly than is usually done. The upshot is, however, an even stronger vindication of the experimental method and an even deeper challenge to the received wisdom.

Laboratory experiments are a means of controlling the social environment so that experiments can be replicated and the results from different experiments can be compared. In physics and chemistry, the experimental method has the additional goal of *eliminating all influences on the behavior of the object of study except those controlled by the experimenter*. This goal can be achieved because elementary particles, and even chemical compounds, are completely interchangeable, given a few easily measurable characteristics (atomic number, energy, spin, chemical composition, and the like). Experiments in human social interaction, however, *cannot* achieve this goal, even in principle, because experimental subjects bring their personal history with them into the laboratory. Their behavior is therefore *ineluctably* an interaction between the subject's personal history and the experimenter's controlled laboratory conditions.

This observation is intimately related to the basic structure of evolutionary game theory (as well as human psychology, as stressed by Loewenstein 1999). As we have seen (§5.17), in strategic interaction nature abhors low probability events, and for an experimental subject, *the experiment is precisely a low probability event!* Neither personal history nor general cultural/genetic evolutionary history has prepared subjects for the Ultimatum, Dictator, Common Pool Resource, and other games that they are asked to confront. As we have suggested in §5.17, an agent treats a low probability event as a high probability event by assigning a novel situation to one of a small number of pre-given *situational contexts*, and then deploying the behavioral repertoire—payoffs, probabilities, and actions—appropriate to that context. We may call this *choosing a frame* for interpreting the experimental situation. This is how subjects bring their history to an experiment.[3]

The results of the ultimatum game (§11.4.1), for instance, suggest that in a two-person bargaining situation, in the absence of other cues, the situational context applied by most subjects dictates some form of "sharing." Suppose we change the rules such that both proposer and respondent are members of different *teams* and each is told that their respective winning will be paid to the team rather than the individual. A distinct situational context, involving "winning," is now often deemed appropriate, dictating acting on behalf of one's team and suppressing behaviors that would be otherwise individually satisfying—such as "sharing." In this case, proposers offer much less, and respondents very rarely reject positive offers (Shogren 1989). Similarly, if the experimenters introduce notions of property rights into the strategic situation (e.g., that the proposer in an ultimatum game has "earned" or "won" the right to this position), then motivations concerning "fairness" are considerably attenuated in the experimental results (Hoffman, McCabe, Shachat, and Smith 1994, Hoffman, McCabe, and Smith 1996).

In short, laboratory experiments (a) elucidate how subjects identify situational contexts, and then (b) describe how agents react to the formal parameters and material payoffs, subject to the situational contexts they have identified.

[3] For a similar view, see Hoffman, McCabe, and Smith (1996). A caveat: It is incorrect to think that the subjects are "irrational" or "confused" because they drag their history into an experimental situation. In fact, they are acting normally on the basis of the preferences they exhibit in daily life. Of course, if this low probability event (being a subject in an experiment) turns into a high probability event (e.g., by being repeatedly asked to be a subject), agents may change their framing or even create a wholly new situational context for the purpose at hand. The process is not well understood.

11.2.2 Self-Interest and Rationality

The culture surrounding economics as a discipline fosters the belief that rationality implies self-interest, outcome-orientation, and time-consistency.[4] No such implication can be supported. A *rational agent* draws conclusions logically from given premises, has premises that are defensible by reasoned argument, and uses evidence dispassionately in evaluating factual assertions. This is reflected in economic theory by a rational agent having transitive preferences and maximizing an appropriate objective function over an appropriate choice set (§11.3.2). Since rationality does not presuppose unlimited informational processing capacities and perfect knowledge, even Herbert Simon's (1982) concept of *bounded rationality* is consistent with a fully rational agent.[5]

I have never seen a serious argument supporting the assertion that rationality in either the everyday sense, or in the narrower sense of optimizing subject to constraints, implies self-interest, outcome-orientation, or time-consistency. Perhaps Milton Friedman's (1953) suggestion that assumptions are justified by the conclusions they support is the most worthy of notice. But since the *Homo economicus* model does not predict well outside of impersonal market situations, his argument is no help here. The most common "informal" argument is reminiscent of Louis XIV's après moi le déluge defense of the monarchy: drop these assumptions and we lose the ability to predict altogether. The models developed recently in the professional literature, some of which are presented below, show that we have little to fear from the Flood.

In neither the everyday nor the narrower economic sense of the term does rationality imply self-interest. It is just as "rational" for me to prefer to have *you* enjoy a fine meal as for me to enjoy the meal myself. It is just as "rational" for me to care about the rain forests as to care about my beautiful cashmere sweater. And it is just as "rational" to reject an unfair offer as it is to discard an ugly article of clothing.

Evolutionary game theory treats agents' objectives as a matter of fact, not logic, with a presumption that these objectives must be compatible with

[4] I use the term "self-interested" to mean *self-regarding*. Self-regarding agents evaluate alternative states of the world by considering only their impact on themselves, narrowly construed.

[5] Indeed, it can be shown (Zambrano 1997) that every boundedly rational agent is a fully rational agent subject to an appropriate set of Bayesian priors concerning the state of nature.

an appropriate evolutionary dynamic. We can just as well build models of regret, altruism, vindictiveness, status-seeking, and addiction as of choosing a bundle of consumption goods subject to a budget constraint (Gintis 1972a, 1972b, 1974, 1975, Bowles and Gintis 1993b, Becker and Murphy 1988, Becker 1996, Becker and Mulligan 1997). As suggested below, evolutionary models do not predict self-interested behavior.

Far from being the norm, people who are self-interested are in common parlance called *sociopaths*. A sociopath treats others instrumentally, either without regard for their feelings (e.g., a sexual predator, a cannibal, or a professional killer), or evaluates the feelings of others only according to their effect on the sociopath (e.g., a sadist or a bully). A neoclassical economist may respond that the postulate of self-interest applies only to the market phenomena that economists normally study—even Adam Smith, the architect of the *invisible hand*, was also the author of the Theory of Moral Sentiments, according to which the principle of *sympathy* guides the social relations among people. But social interactions, even in economics, are not restricted to impersonal market contexts. Moreover, by deploying the appropriate game-theoretic models (§11.5), we shall see that we can predict when non-self-interested agents will behave in a self-interested manner. We need not assume self-interest from the outset.

11.3 Behavioral Economics: Games against Nature and against Ourselves

Problems with the *Homo economicus* model arise even prior to strategic interactions among multiple agents, with *games against Nature* and *games against ourselves*.[6] A "game against nature" involves a single agent choosing under conditions of uncertainty, where none of the uncertainty is strategic—that is, either the uncertainty is due to natural acts (crop loss, death, and the like) or, if other people are involved, the others do not behave strategically toward the agent being modeled. A "game against oneself" is a choice

[6] For pedagogical reasons I am giving this experimental material the benefit of the doubt. Many of the situations subjects face in behavioral experiments are precisely the "zero probability events" for which we expect individuals to be evolutionarily unprepared. Thus, as Ken Binmore (1999) stresses, we cannot presume that because human subjects do not find the optima in complex decision problems in the laboratory, they are incapable of finding the optima in the everyday-life situations that use the same analytical principles. Indeed, under appropriate conditions we would expect roughly optimal solutions to replicate and diffuse in a population without individuals ever having to "solve" the underlying problems.

situation in which an agent optimizes over time, but cannot automatically precommit to carrying out in the future the plans being made in the present. Experimental evidence supports the following generalizations concerning such games:

- People appear to have higher discount rates over payoffs in the near future than in the distant future. It follows that people often favor short term gains that entail long term losses. We often term this "impulsivity" or "weakness of will." Technically this is called *hyperbolic discounting*.
- People often have inconsistent preferences over lotteries and do not know and cannot apply the laws of probability consistently without extensive formal training. Rather, people use informal heuristics and socially acquired rules to choose among risky alternatives.
- The *Homo economicus* model assumes that people react to the absolute level of payoffs, whereas in fact they tend to privilege the status quo (their current position) and are sensitive to changes from the status quo. In particular, people tend to exhibit *loss aversion*, making them risk-loving over losses and risk-averse over gains at the same time (§16.41).

11.3.1 Time Inconsistency and Hyperbolic Discounting

"Time consistency" means that the future actions required to maximize the current present value of utility remain optimal in the periods when the actions are to be taken.[7] The central theorem on choice over time is that time consistency requires that utility be additive and independent across time periods, with future utilities discounted to the present at a fixed rate (Strotz

[7] I do not know why it is considered "rational" to be time consistent. There are no plausible models within which time consistency has optimal welfare-enhancing properties. In a strategic setting, time consistency can entail lower payoffs. For instance, if I cannot control my temper (getting angry today has a higher value than paying the costs tomorrow), and you know this, you may give me my way, whereas if I were time consistent, you would know that I won't actually blow a fuse.

Of course, if you are not time consistent, and if you know this fact, you will not commit yourself to future choices that you know you will not make, or you will precommit yourself to making certain future choices, even at a cost. For instance, if you are saving in year 1 for a purchase in year 3, but you know you will be tempted to spend the money in year 2, you can put it in a bank account that cannot be accessed until the year after next. My former teacher Leo Hurwicz called this the "piggy bank effect."

1955).[8] Are people time consistent? Take, for instance, impulsive behavior. Economists are wont to argue that what appears to be "impulsive"—cigarette smoking, drug use, unsafe sex, overeating, dropping out of school, punching out your boss, and the like—may in fact be welfare-maximizing for people who have high time discount rates or who prefer acts with high future costs. Controlled experiments in the laboratory cast doubt on this explanation, indicating that people exhibit a systematic tendency to discount the near future at a higher rate than the distant future (Chung and Herrnstein 1967, Loewenstein and Prelec 1992, Herrnstein and Prelec 1992, Fehr and Zych 1994, Kirby and Herrnstein 1995). In fact, observed intertemporal choice appears to fit the model of *hyperbolic discounting* (Ainslie and Haslam 1992, Ainslie 1975, Laibson 1997), first observed by Richard Herrnstein in studying animal behavior (Laibson and Rachlin 1997). In addition, agents have different rates of discount for different types of outcomes (Loewenstein 1987, Loewenstein and Sicherman 1991).

We should not ask why these anomalies occur, because there is no reason to expect time consistency in human behavior in the first place. Since humans appear to share hyperbolic discounting with other species, there is probably an evolutionary explanation for the phenomenon. Impulsiveness may reduce fitness and personal welfare under contemporary environmental circumstances, but it may have contributed to success, or at least imposed little harm, in the conditions under which *Homo sapiens* evolved (Cosmides and Tooby 1992b).

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

[8] Actually, as Lones Smith has pointed out (personal communication), this result assumes that discounting at time t of consumption that occurs at time $s > t$ depends only on the horizon length $s - t$. In general, time consistency does not imply either additivity or a constant discount rate. For instance, aging implies that the probability of death increases with time, so people should apply higher discount rates to the future than to the present (comedian George Burns once exclaimed, when in his nineties, that he never buys green bananas). This behavior does not necessarily imply time inconsistency.

11.3.2 Choice under Uncertainty

The centerpiece of the theory of choice under uncertainty is the *expected utility principle*, which says that "rational" agents choose among lotteries to maximize the expected utility of the payoffs (§3.14, Kreps 1988). Von Neumann and Morgenstern (1944), Friedman and Savage (1948), Savage (1954), and Anscombe and Aumann (1963) showed that the expected utility principle can be derived from the assumption that agents have consistent preferences over an appropriate set of lotteries. By "consistency" we mean *transitivity* and *independence from irrelevant alternatives*, plus some plausible technical conditions. We say preferences are *transitive* if, whenever A is preferred to B, and B is preferred to C, then A is preferred to C.

Why should agents have transitive preferences? An agent who optimizes subject to constraints will certainly have transitive preferences. But the expected utility principle models agents *as if* they were optimizing, when in fact they are *not*, just as we model expert billiards players by assuming that they solve certain systems of differential equations, which of course they do not. If choices are easily reversed, one could fool a nontransitive agent with the "money pump": sell him A, then induce him to pay for B, which he prefers to A, then induce him him to pay for C, which he prefers to B, then induce him to pay for A again. Such an agent fits any plausible definition of "irrational." But lots of choices are *not* reversible, so it is not obviously irrational to be intransitive over such choices.[9]

Independence from irrelevant alternatives is some variation of the following (the particulars depend on the model): suppose two lotteries l_1 and l_2 have the same outcomes with the same probabilities, except for one particular outcome, where the l_1 outcome is preferred to the l_2 outcome. Then lottery l_1 is preferred to lottery l_2. Another version of this is Leonard Savage's *sure-thing principle*, which states that if you prefer l_1 to l_2 when state of the world A occurs, and you also prefer l_1 to l_2 when A does *not* occur, then you prefer l_1 to l_2.[10]

Laboratory testing of the *Homo economicus* model of choice under uncertainty was initiated by the psychologists Daniel Kahneman and Amos

[9] One plausible theory of intransitivity with some empirical support is *regret theory* (Loomes 1988, Sugden 1993).

[10] It is plausible that a rational agent would subscribe to the sure-thing principle. However, people systematically violate the principle because they do not reason well logically over disjunctions without extensive training (Shafir and Tversky 1992, Tversky and Shafir 1992, Shafir 1994).

Tversky. In a famous article in the journal *Science*, Tversky and Kahneman (1974) summarized their early research as follows:

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

Subsequent research has strongly supported this assessment (Kahneman, Slovic, and Tversky 1982, Shafir and Tversky 1992, Shafir and Tversky 1995). Although we still do not have adequate models of these heuristics, we do know the following:

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

b. A second heuristic is that in assessing the frequency of an event, people take excessive account of information that is easily *available* or highly *salient*, even though a selective bias is obviously involved. For this reason, people tend to overestimate the probability of rare events, since such events are highly newsworthy while nonoccurrences are not reported.

c. A third heuristic in problem solving is to start from an initial guess, chosen for its representativeness or salience, and adjust upward or downward toward a final figure. This is called *anchoring* because there is a tendency to underadjust, so the result is too close to the initial guess.

d. Probably as a result of anchoring, people tend to overestimate the probability of conjunctions (p and q) and underestimate the probability of disjunctions (p or q). For an instance of the former, a person who knows an event occurs with 95% probability will overestimate the probability that the event occurs ten times in a row. The actual probability is about 60%. In this case the individual starts with 95% and does not adjust downward sufficiently. Similarly, if a daily event has a failure one time in a thousand, people will underestimate the probability that a failure occurs at least once in a year. The actual probability is 30.5%. Again, the individual starts with 0.1% and doesn't adjust upward enough.

e. People prefer objective probability distributions to subjective distributions derived from applying probabilistic principles, such as the Principle of Insufficient Reason (§16.28.1), which says that if you are completely ignorant as to which of several outcomes will occur, you should treat them as equally probable. For example, if you give a subject a prize for drawing a red ball from an urn containing red and white balls, the subject will pay to have the urn contain 50% red balls rather than contain an indeterminate percentage of red balls.

Choice theorists often express dismay over the failure of people to apply the laws of probability and conform to the axioms of choice theory. This is a strange reaction. People are doubtless applying rules that serve them well in daily life. It takes many years of study to feel at home with the laws of probability, the understanding of which is the product of the last couple of hundred years of scientific research. Moreover, it is costly, in terms of time and effort, to apply these laws even if we know them. Of course, if the stakes are high enough, it is worthwhile to go to the effort, or engage an expert who will do it for you. But generally, as Kahneman and Tversky suggest, we apply a set of heuristics that more or less get the job done. Among the most prominent heuristics is simply *imitation:* decide what class of phenomenon is involved, find out what people "normally do" in that situation, and do it. If there is some mechanism leading to the survival and growth of relatively successful behaviors (see chapter 9), and if the problem in question recurs with sufficient regularity, the choice-theoretic solution will describe the winner of a dynamic social process of trial, error, and imitation.

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

11.3.3 Loss Aversion and Status Quo Bias

It appears that people value gains and losses not according to their absolute levels, but rather according to their deviation from their current position (Helson 1964). The most venerable expression of this principle is the *Weber-Fechner Law* that initiated the science of psychophysics. According to this law, a just-noticeable change in a stimulus is a fixed ratio of the level of the stimulus. The assumption that utility functions are concave, and hence individuals are risk-averse, is of course based on similar reasoning, as is the notion that individuals "adjust" to an accustomed level of income (§16.41), so that subjective well-being is associated more with *changes* in income rather than with the *level* of income. See, for instance Easterlin (1974, 1995), Lane (1991, 1993), and Oswald (1997).

Experimental evidence supports an even stronger assertion: *people are about twice as averse to taking losses as to enjoying an equal level of gains* (Kahneman, Knetch, and Thaler 1990, Tversky and Kahneman 1981b). This means, for instance, that an individual may attach zero value to a lottery that offers an equal chance of winning $1000 and losing $500. This also implies people are *risk-loving over losses*, while they remain risk-averse over gains. For instance, many individuals will choose a 25% probability of losing $2000 rather than a 50% chance of losing $1000 (both have the same expected value, of course, but the former is riskier).

One implication of loss aversion is the *endowment effect* (Kahneman, Knetch, and Thaler 1991), according to which people place a higher value on what they possess than they place on the same things when they do not possess them. For instance, if you win a bottle of wine that you could sell for $200, you may drink it rather than sell it, but you would never think of buying even a $100 bottle of wine. Not only does the endowment effect exist, but there is evidence that people underestimate it and hence cannot rationally correct for it in their choice behavior (Loewenstein and Adler 1995).

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

Yet another implication is a *status quo bias*, according to which people often prefer the status quo over any of the alternatives, but if one of the alternatives becomes the status quo, that too is preferred to any of the alternatives (Kahneman, Knetch, and Thaler 1991). The status quo makes sense if we recognize that any change can involve a loss, and since on the average gains do not offset losses, it is possible that any one of a number of alternatives might be preferred if it is the status quo.

11.4 Experimental Game Theory: The Laboratory Meets Strategic Interaction

Many of the anomalies in testing game-theoretic predictions of strategic interaction flow from two observations:

- The *Homo economicus* model assumes preferences are *self-regarding* and *outcome-regarding*, whereas preferences are also *other-regarding* and *process-regarding.* In particular, people care about fairness (§11.5), reciprocity (§11.6), and group membership (§11.9).
- The *Homo economicus* model assumes preferences are *exogenous:* they are determined outside of, and substantially unaffected by, the structure of strategic interaction or any other substantive aspect of the economy. However, preferences are partly *endogenous*, depending both on the agent's personal history and the nature of the strategic interaction in which the agent is engaged. In particular, by *choosing a frame* (§11.2.1), an agent chooses to act according to a particular pattern of subjective payoffs.

As a basis for interpreting a broad range of experiments, I will introduce several new *personas*, the most novel of whom I call *Homo reciprocans*. *Homo reciprocans*' behavior in market situations, in which punishing and rewarding are impossible or excessively costly, is much like that of *Homo economicus*. But *Homo reciprocans* comes to strategic interactions with a propensity to cooperate, responds to cooperative behavior by maintaining or increasing his level of cooperation, and responds to noncooperative behavior by retaliating against the "offenders," even at a cost to himself, and even when he could not reasonably expect future personal gains to flow from such retaliation. When other forms of punishment are not available, *Homo reciprocans* responds to defection with defection, leading to a downward spiral of noncooperation. *Homo reciprocans* is thus neither the selfless altruist of utopian theory, nor the selfish hedonist of neoclassical economics. Rather, he is a conditional cooperator whose penchant for reciprocity can be elicited under circumstances in which personal self-interest would dictate otherwise.[11] A second, probably more familiar, persona is *Homo egualis*, who cares not only about his own payoff, but also about how it compares with the payoff of others. *Homo egualis* may be willing to reduce his own payoff to increase the degree of equality in the group (whence widespread support for charity and social welfare programs). But he is especially displeased when subjected to *relative deprivation*, by being placed on the losing end of an unequal relationship. Indeed, *Homo egualis* may be willing to reduce his own payoff if that reduces the payoff of relatively favored players even more (Loewenstein, Thompson, and Bazerman 1989).

A third and also quite familiar persona is *Homo parochius*, who divides the world into *insiders* and *outsiders* according to context-dependent and even apparently arbitrary characteristics, values insiders' welfare more highly than that of outsiders, evaluates insiders' personal qualities more highly than that of outsiders, and partially suppresses personal goals in favor of the goals of the group of insiders. Race, ethnicity, common language, and nationality are well-known examples of characteristics that are used to distinguish "insiders" from "outsiders." But in experimental settings, subjects exhibit parochial preferences even when the basis of group membership is explicitly random or arbitrary, and there is no a priori reason for subjects to care about others in their own as opposed to other groups (Turner 1984).

11.4.1 The Ultimatum Game

The *ultimatum game*, first studied by Werner Güth, Rolf Schmittberger, and Berndt Schwarze (1982), is a showcase for costly retaliation in a one-shot

[11] Another aspect of reciprocity is commonly known as "gift exchange," in which one agent behaves more kindly than required toward another, with the hope and expectation that the other will respond kindly as well (Akerlof 1982). For instance, in an laboratory-simulated work situation in which "employers" can pay higher than market-clearing wages in hopes that "workers" will reciprocate by supplying a high level of effort, see Fehr, Gächter, Kirchler, and Weichbold 1998 and Fehr, Gächter, and Kirchsteiger 1997.

situation. Under conditions of anonymity, one player, called the "proposer," is handed a sum of money, say $10, and is told to offer any number of dollars, from $1 to $10, to the second player, who is called the "responder." The responder, again under conditions of anonymity, can either accept the offer or reject it. If the responder accepts the offer, the money is shared accordingly. If the responder rejects the offer, both players receive nothing.

There are *lots* of Nash equilibria in this game. In fact, there are $2^{10} - 1 = 1023$ Nash equilibria, since every pattern of "accept/reject" of the ten numbers $1, \ldots, 10$, except "reject every offer" is part of a Nash equilibrium, and the best response for the proposer to such a strategy is the smallest number in that pattern that the responder will accept ("reject every offer" is never a best response). For instance, one Nash equilibrium is for the responder to accept (3,7,10) and reject all other offers, and for the proposer to offer 3.

But, there is only *one* responder strategy that is subgame perfect: accept anything you are offered. However, when actually played by people, *the subgame perfect outcome is almost never attained or even approximated.* In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, proposers routinely offer respondents very substantial amounts (50% of the total being the modal offer), and respondents frequently reject offers below 30% (Camerer and Thaler 1995, Güth and Tietz 1990, Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991). These results are obtained in experiments with stakes as high as three months' earnings.[12]

Are these results culturally dependent? Do they have a strong genetic component, or do all "successful" cultures transmit similar values of reciprocity to individuals? Roth et al. (1991) conducted ultimatum games in four different countries (United States, Yugoslavia, Japan, and Israel) and found that while the level of offers differed a bit in different countries, the probability of an offer being rejected did not. This indicates that both proposers and respondents share the same notion of what is considered "fair" in that society, and that proposers adjust their offers to reflect this common notion. The differences in level of offers across countries, by the way, were relatively small.

[12] For analyses of ultimatum games, see Forsythe, Horowitz, Savin, and Sefton 1994, Hoffman, McCabe, Shachat, and Smith 1994, Hoffman, McCabe, and Smith 1998, Cameron 1995, and Fehr and Tougareva 1995.

By contrast, Henrich (2000) carried out ultimatum games among the Machiguenga Indians, a Peruvian Amazon hunter-gatherer tribe, and found significant differences. Among the Machiguenga, average offers were much lower, the median offer being about 20%, offers of 15% were often accepted, and the rejection rate was very low. These anthropological results are currently being extended by studying additional societies in various parts of the world (Bowles, Boyd, Fehr, and Gintis 1997).

In the United States and other complex societies, when asked why they offer more than the lowest possible amount, proposers commonly say that they are afraid that respondents will consider low offers unfair and reject them. When respondents reject offers, they give virtually the same reasons for their actions.[13]

11.4.2 The Public Goods Game

Another important experimental setting in which strong reciprocity has been observed is that of the *public goods game,* designed to illuminate such problems as the voluntary payment of taxes and contribution to team and community goals. Public goods experiments have been run many times, under varying conditions, beginning with the pioneering work of the sociologist G. Marwell, the psychologist R. Dawes, the political scientist J. Orbell, and the economists R. Isaac and J. Walker in the late 1970s and early 1980s.[14] The following is a common variant of the game. Ten subjects are told that $1 will be deposited in each of their "private accounts" as a reward for participating in each round of the experiment. For every $1 that a subject moves from his "private account" to the "public account," the experimenter will deposit one half dollar in the private accounts of each of the subjects at the end of the game. This process will be repeated ten times, and at the end the subjects can take home whatever they have in their private accounts.

If all ten subjects are perfectly cooperative, each puts $1 in the public account at the end of each round, generating a public pool of $10; the experimenter then puts $5 in the private account of each subject. After ten rounds of this, each subject has $50. Suppose, by contrast, that one subject is perfectly selfish, while the others are cooperative. The selfish one keeps his

[13] In all of the above experiments a significant fraction of subjects (about a quarter, typically) conform to the self-interested preferences of *Homo economicus*, and it is often the self-serving behavior of this minority that, when it goes unpunished, unravels initial generosity and cooperation when the game is repeated.

[14] For a summary of this research and an extensive bibliography, see Ledyard (1995).

$1-per-round in his private account, whereas the cooperative ones continue to put $1 in the public pool. In this case, the selfish subject who takes a free ride on the cooperative contributions of others ends up with $55 at the end of the game, while the other players will end up with $45 each. But if all players opt for the selfish payoff, then no one contributes to the public pool, and each ends up with $10 at the end of the game. And if one player cooperates, while the others are all selfish, that player will end up with $5 at the end of the game while the others will get $15. It is thus clear that this is indeed an "iterated Prisoner's Dilemma"—whatever other players do on a particular round, a player's highest payoff comes from contributing nothing to the public account. If others cooperate, it is best to take a free ride; if others are selfish, it is best to join them. But if no one contributes, all receive less than they would if all had cooperated.

Public goods experiments show that only a fraction of subjects conform to the *Homo economicus* model, contributing nothing to the public account. Rather, in a one-stage public goods game, people contribute on average about half of their private account. The results in the early stages of a repeated public goods game are similar. In the middle stages of the repeated game, however, contributions begin to decay, until at the end they are close to the *Homo economicus* level—i.e., zero.

Could we not explain the decay of public contribution by *learning*: the participants really do not understand the game at first, but once they hit upon the free-riding strategy, they apply it? Not at all. One indication that learning does not account for the decay of cooperation is that increasing the number of rounds of play (when this is known to the players) leads to a decline in the rate of decay of cooperation (Isaac, Walker, and Williams 1994). Similarly, Andreoni (1988) finds that when the whole process is repeated with the same subjects but with different group composition, the initial levels of cooperation are restored, but once again cooperation decays as the game progresses. Andreoni (1995) suggests a *Homo reciprocans* explanation for the decay of cooperation: public-spirited contributors want to retaliate against free-riders, and the only way available to them in the game is by not contributing themselves.

11.4.3 The Public Goods Game with Retaliation

Could the decay of cooperation in the public goods game be due to cooperators retaliating against free-riders by free-riding themselves? Subjects often report this behavior retrospectively. More compelling, however, is the fact that when subjects are given a more constructive way of punishing defectors, they use it in a way that helps sustain cooperation (Dawes, Orbell, and Van de Kragt 1986, Sato 1987, Yamagishi 1988a, 1988b, 1992).

For instance, in Ostrom, Walker, and Gardner (1992) subjects interacted for about twenty-five periods in a public goods game, and by paying a "fee," subjects could impose costs on other subjects by "fining" them. Since fining costs the individual who uses it but the benefits of increased compliance accrue to the group as a whole, the only subgame perfect Nash equilibrium in this game is for no player to pay the fee, so no player is ever punished for defecting, and all players defect by contributing nothing to the public account. However, the authors found a significant level of punishing behavior. The experiment was then repeated with subjects being allowed to communicate, without being able to make binding agreements. In the framework of the *Homo economicus* model, such communication is called *cheap talk* and cannot lead to a distinct subgame perfect equilibrium. But in fact such communication led to almost perfect cooperation (93%) with very little sanctioning (4%).

The design of the Ostrom-Walker-Gardner study allowed individuals to engage in strategic behavior, since costly retaliation against defectors could increase cooperation in future periods, yielding a positive net return for the retaliator. It is true that backward induction rules out such a strategy, but we know that people do not backward induct very far anyway. What happens if we remove any possibility of retaliation being strategic? This is exactly whatFehr and Gächter (1999) studied. They set up a repeated public goods game with the possibility of costly retaliation, but they ensured that group composition changed *in every period* so subjects knew that costly retaliation could not confer any pecuniary benefit to those who punish. Nonetheless, punishment of free-riding was prevalent and gave rise to a large and sustainable increase in cooperation levels.

11.4.4 The Common Pool Resource Game

In 1968 Garrett Hardin wrote a famous article in the journal *Science* entitled "The Tragedy of the Commons" (Hardin 1968). The term "commons" referred originally to the region of an English village that belonged to the villagers as a group and on which villagers were permitted to graze their sheep or cows. The "tragedy" in the tragedy of the commons was that the commons tended to be overgrazed, since each villager would graze to the point where the *private* costs equal the benefits, whereas grazing imposed additional *social* costs on the rest of the community. We explored this phenomenon rather fancifully in Klingon and Snarks (§3.9), but in fact it applies to what are termed *common pool resources* in general. Some involve social problems of the highest importance, including air and water pollution, overfishing, overuse of antibiotics, traffic congestion, excessive groundwater use, overpopulation, and the like.

The general implication from Hardin's analysis was that some centralized entity, such as a national government or international agency, had to step in to prevent the tragedy by regulating the common. The historical experience in regulating the commons, however, has been a patchwork of successes and failures. In 1990 Elinor Ostrom published an influential book, *Governing the Commons*, suggesting that the Hardin analysis did not apply generally, since local communities often had ways of self-organizing and self-governing to prevent overexploitation of the commons, and that government policy often exacerbated rather than ameliorated the problem by undermining the social connections on which local regulation was based.

When formalized as a game, the common pool resource problem is simply an n-person repeated Prisoner's Dilemma, in which each player hopes the other players will cooperate (not take too much of the common resource), but will defect (take too much) no matter what the other players do. Since the public goods game (§11.4.2) is also an n-person repeated prisoner's dilemma, it is not surprising that both in the real world and in experimental settings, under the appropriate conditions, we see much more cooperation than predicted by the *Homo economicus* model.

Ostrom, Gardner, and Walker (1994) used both experimental and field data to test game-theoretic models of common pool resources. They found more spontaneous cooperation in the field studies than predicted, and when communication and sanctioning were permitted in the laboratory, the level of cooperation became quite high.

While common pool resource and public goods games are equivalent for *Homo economicus,* people treat them quite differently in practice. This is because the status quo in the public goods game is the individual keeping all the money in the private account, while the status quo in the common pool resource game is that the resource is not being used at all. This is a good example of a *framing effect* (§11.3.3), since people measure movements from the status quo and hence tend to undercontribute in the public goods game and overcontribute (underexploit) in the common pool resource game, compared to the social optimum (Ostrom 1998).

In the real world, of course, communities often do *not* manage their common pool resources well. The point of Ostrom's work is to identify the sources of failure, not to romanticize small communities and informal organization. Among other reasons, the management of common pool resources fails when communities are so large that it pays to form a local coalition operating against the whole community, and when resources are so unequally distributed that it pays the wealthy to defect on the nonwealthy and conversely (Hackett, Schlager, and Walker 1994, Bardhan, Bowles, and Gintis 2000).

11.5 *Homo egualis*

Homo egualis exhibits a *weak* urge to reduce inequality when on top, and a *strong* urge to reduce inequality when on the bottom. Since the advent of hierarchical societies that are based on settled agriculture, societies have attempted to inculcate in its less fortunate members precisely the opposite values—subservience to and acceptance of the status quo. The widely observed distaste for relative deprivation is thus probably a genetically based behavioral characteristic of humans. Since small children spontaneously share (even the most sophisticated of primates, such as chimpanzees, fail to do this), the urge of the fortunate to redistribute may also be part of human nature, though doubtless a weaker impulse in most of us.

Support for *Homo egualis* comes from the anthropological literature. *Homo sapiens* evolved in small hunter-gatherer groups. Contemporary groups of this type, although widely dispersed throughout the world, display many common characteristics. This commonality probably reflects their common material conditions. From this and other considerations we may tentatively infer the social organization of early human society from that of these contemporary foraging societies.[15]

Such societies have no centralized structure of governance (state, judicial system, church, Big Man), so the enforcement of norms depends on the voluntary participation of peers. There are many unrelated individuals, so

[15] See Woodburn 1982, Boehm 1982, 1993, Blurton-Jones 1987, Cashdan 1980, Knauft 1991, Hawkes 1992, 1993, ,, Kaplan and Hill 1985a,b, , Kaplan, Hill, Hawkes, and Hurtado 1984, Lee 1979, Woodburn and Barnard 1988, Endicott 1988, Balikci 1970, Kent 1989, Damas 1972, Wenzel 1995, Knauft 1989.

cooperation cannot be explained by kinship ties. Status differences are very circumscribed, monogamy is widely enforced,[16] members who attempt to acquire personal power are banished or killed, and there is widespread sharing of large game and other food sources that are subject to substantial stochasticity, independent of the skill and/or luck of the hunters. Such conditions are, of course, conducive to the emergence of *Homo egualis*.

We model *Homo egualis* following Fehr and Schmidt (1999). Suppose the monetary payoffs to n players are given by $x = (x_1, \ldots, x_n)$. We take the utility function of player i to be

$$u_i(x) = x_i - \frac{\alpha_i}{n-1}\sum_{x_j > x_i}(x_j - x_i) - \frac{\beta_i}{n-1}\sum_{x_j < x_i}(x_i - x_j). \qquad (11.1)$$

A reasonable range of values for β_i is $0 \leq \beta_i < 1$. Note that when $n = 2$ and $x_i > x_j$, if $\beta_i = 0.5$ then i is willing to transfer income to j dollar for dollar until $x_i = x_j$, and if $\beta_i = 1$ then i is willing to throw away money until $x_i = x_j$. We also assume $\beta_i < \alpha_i$, reflecting the fact that people are more sensitive to inequality when on the bottom than when on the top.

We shall show that with these preferences, we can reproduce some of the salient behaviors in ultimatum and public goods games, where fairness appears to matter, as well as in market games where it does not.

Consider first the ultimatum game. Let y be the share the proposer offers the respondent, so the proposer gets $x = 1 - y$. Since $n = 2$, we can write the two utility functions as

$$u(x) = \begin{cases} x - \alpha_1(1-2x) & x \leq 0.5 \\ x - \beta_1(2x-1) & x > 0.5 \end{cases} \qquad (11.2)$$

$$v(y) = \begin{cases} y - \alpha_2(1-2y) & y \leq 0.5 \\ y - \beta_2(2y-1) & y > 0.5 \end{cases} \qquad (11.3)$$

We have the following theorem.

THEOREM 11.1 *Suppose the payoffs in the ultimatum game are given by (11.2) and (11.3), and* α_2 *is uniformly distributed on the interval* $[0, \alpha^*]$. *Writing* $y^* = \alpha^*/(1 + 2\alpha^*)$, *we have the following:*

a. If $\beta_1 > 0.5$ *the proposer offers* $y = 0.5$.

[16] Monogamy in considered to be an extremely egalitarian institution for men, since it ensures that virtually all adult males will have a wife.

b. If $\beta_1 = 0.5$ the proposer offers $y \in [y^*, 0.5]$.
c. If $\beta_1 < 0.5$ the proposer offers y^*.

In all cases the respondent accepts. We leave the proof, which is straightforward, to the reader.

Now suppose we have a public goods game $\mathcal{G}$ with $n \geq 2$ players. Each player i is given an amount 1 and decides independently what share x_i to contribute to the public account, after which the public account is multiplied by a number a with $1 > a > 1/n$ and shared equally among the players. The monetary payoff for each player then becomes $1 - x_i + a\sum_{j=1}^{n}, x_j$ and the utility payoffs are given by (11.1). We then have this theorem.

THEOREM 11.2 *In the n-player public goods game $\mathcal{G}$,*

a. If $\beta_i < 1-a$ for player i, then contributing nothing to the public account is a dominant strategy for i.
b. If there are $k > a(n-1)/2$ players with $\beta_i < 1-a$, then the only Nash equilibrium is for all players to contribute nothing to the public account.
c. If there are $k < a(n-1)/2$ players with $\beta_i < 1-a$ and if all players i with $\beta_i > 1-a$ satisfy $k/(n-1) < (a+\beta_i-1)/(\alpha_i+\beta_i)$, then there is a Nash equilibrium in which the latter players contribute all their money to the public account.

Note that if a player has a high β and hence could possibly contribute, but also has a high α so the player strongly dislikes being below the mean, then condition $k/(n-1) < (a+\beta_i-1)/(\alpha_i+\beta_i)$ in part (c) of the theorem will fail. In other words, cooperation with defectors requires that contributors not be excessively sensitive to relative deprivation.

The proof of the theorem is a bit tedious but straightforward, and will be left to the reader. We prove only (c). We know from (a) that players i with $\beta_i < 1-a$ will not contribute. Suppose $b_i > 1-a$, and assume all other players satisfying this inequality contribute all their money to the public account. By reducing his contribution by $\delta > 0$, player i saves $(1-a)\delta$ directly plus receives $k\alpha_i\delta/(n-1)$ in utility from the higher returns compared to the noncontributors, minus $(n-k-1)\delta\beta_i$ in utility from the lower returns compared with the contributors. The sum must be nonpositive in a Nash equilibrium, which reduces to the inequality in (c).

Despite the fact that players have egalitarian preferences given by (11.1) if the game played has sufficiently marketlike qualities, the unique Nash equilibrium may settle on the competitive equilibrium, however "unfair" this appears to be to the participants. Consider the following.

THEOREM 11.3 *Suppose preferences are given by (11.1) and $1 is to be shared between player 1 and one of the players $i = 2, \ldots, n$, who submit simultaneous bids y_i for the share they are willing to give to player 1. The highest bid wins, and among equal highest bids, the winner is drawn at random. Then, for any set of (α_i, β_i), in every subgame perfect Nash equilibrium player 1 receives the whole $1.*

The proof is left to the reader. Show that at least two bidders will set their $y_i's$ to 1, and the seller will accept this offer.

11.6 *Homo reciprocans:* Modeling Strong Reciprocity

While models of *Homo egualis* can explain many experimental results that are anomalous in terms of *Homo economicus*, other experiments suggest that agents care about the *intentions* of their partners as well as the distributional outcomes. For instance, if offers in an ultimatum game are generated by a computer rather than proposers, and if respondents knows this, low offers are much less likely to be rejected (Blount 1995). This suggests that players are motivated by *reciprocity*, reacting to a violation of behavioral norms rather than simply seeking a more equitable distribution of outcomes (Greenberg and Frisch 1972).

The importance of reciprocity in strategic interaction is common to many forms of life, as has been stressed by Robert Trivers (1971) in his seminal work on reciprocal altruism. The robustness of reciprocal behavior appears in computer simulations as well, as in the work of Robert Axelrod and W. D. Hamilton (1981) on Tit-for-Tat strategies (§6.11). Artificial life simulations of repeated-interaction prisoner's dilemma games also show the robustness of strategies that are "nice" by never defecting first, "punishing" by always punishing defection, and "forgiving" by returning to cooperation after a short period of punishing, if the other player is cooperating.[17]

[17] Nowak and Sigmund (1992) show that when players can make mistakes, a more forgiving version of Tit-for-Tat, called Generous Tit-for-Tat, drives out Tit-for-Tat. Nowak and Sigmund (1993) show that when players can respond to their own as well as their opponents' moves, strategies that repeat moves when successful and switch when unsuccessful (Pavlov) outcompete Tit-for-Tat under the same simulation conditions as present in the original computer contests run by Axelrod. Laboratory experiments with humans

However, as we know from the theory of repeated games (§6.4), reciprocity in the above sense is just "enlightened self-interest," which is fully compatible with the *Homo economicus* model, as long as discount rates are sufficiently low. In effect, Trivers's reciprocal altruist and the Axelrod-Hamilton Tit-for-Tatter behave little differently from *Homo economicus* with an appropriate time discount rate. The *Homo reciprocans* who emerges from laboratory experiments, by contrast, provides a more robust basis for prosocial behavior, since he does not depend upon frequently repeated interactions to induce him to cooperate and punish defectors.

Homo reciprocans exhibits what may be called *strong reciprocity*, by which we mean a propensity to cooperate and share with others similarly disposed, even at personal cost, and a willingness to punish those who violate cooperative and other social norms, even when punishing is personally costly, and even when there are no plausible future rewards or benefits from so behaving.

Critics of the notion of strong reciprocity suggest that reciprocal behavior in one-shot games is just a confused carryover of the subject's extensive experience with repeated games in everyday life to the rare experience of the one-shot game in the laboratory. This is incorrect. Human beings in contemporary society are engaged in one-shot games with very high frequency—virtually every interaction we have with strangers is of this form. Major rare events in people's lives (fending off an attacker, battling hand-to-hand in wartime, experiencing a natural disaster or major illness) are one-shots in which people appear to exhibit strong reciprocity much as in the laboratory. Moreover, *the fact that humans often "confuse" one-shots and repeated interactions, when they clearly have the cognitive mechanisms to distinguish, suggests that the "confusion" may be fitness-enhancing*. It is therefore misleading to suggest that conflating one-shot and repeated games is a regrettable human weakness.

Consider a two-person extensive form game $\mathcal{G}$, where a fairness norm (a situational context that players may or may not apply to the game) suggests equal payoffs to all players—examples include the Prisoner's Dilemma, the ultimatum game, and the Common Pool Resources game.[18] Let $\pi_i(p_1, p_2)$ be the payoff to player $i = 1, 2$ when i uses behavioral strategy p_i (§4.13), and let $\pi_i(p_1, p_2|\nu)$ be the payoff to i, conditional on being at information

(Wedekind and Milinski 1993) show that human subjects use Generous Tit-for-Tat and Pavlov, as well as more sophisticated Pavlov-like strategies.

[18] This model of strong reciprocity follows Falk and Fischbacher 1998.

set ν. The *fairness* $f_j(p_1, p_2|\nu)$ of j if it is $i \neq j$'s move at ν is defined by

$$f_j(p_1, p_2|\nu) = \pi_i(p_1, p_2|\nu) - \pi_j(p_1, p_2|\nu).$$

Thus, at ν, j has been relatively generous if $f_j > 0$, and relatively selfish if $f_j < 0$.

For every pure action a available to i at ν, let $p_i(a)$ be the behavioral strategy for i that is the same as p_i everywhere except at ν, where i takes action a. We then define i *kindness* from taking action a at ν to be

$$k_i(p_1, p_2, a|\nu) = \pi_j((p_i(a), p_j)|\nu) - \pi_j(p_1, p_2|\nu),$$

where $(p_i(a), p_j) = p_1(a), p_2$ if $i = 1$ and $(p_i(a), p_j) = p_1, p_2(a)$ if $i = 2$. In other words, given the pair of strategies (p_1, p_2), player i who moves at node ν is being "kind" when choosing move a if this gives j a greater payoff than that indicated by p_i.

The total payoff to i at a terminal node $t \in T$ of $\mathcal{G}$ is then

$$u_i(t) = \pi_i(t) + \rho_i \sum_{\nu \in N_i(t)} f_j(p_1, p_2|\nu) k_i(p_1, p_2, a_\nu|\nu), \qquad (11.4)$$

where $N_i(t)$ is the set of information sets where i moves on the path to t, and a_ν is the action at ν on the path to t. Note that if $f_j > 0$ at a certain node, then *ceteris paribus* player i gains from exhibiting positive kindness, while if $f_j < 0$, the opposite is the case. Note also that these payoffs are relative to a specific pair of behavioral strategies (p_1, p_2). This aspect of (11.4) reflects the fact that *Homo reciprocans* cares not only about payoffs, but also about the actions of the other player. We say that a pair of strategies (p_1^*, p_2^*) of $\mathcal{G}$ is a *reciprocity equilibrium* if (p_1^*, p_2^*) is a Nash equilibrium of (11.4) when (p_1, p_2) is replaced by (p_1^*, p_2^*) on the right-hand side of (11.4).[19]

THEOREM 11.4 *Suppose both players in an ultimatum game have preferences given by (11.4), where $\rho_1, \rho_2 > 0$ are known by both players, and let s be the share the proposer offers the respondent. Let $p^*(s)$ be the respondent's best reply to the offer s, and let $(p^*(s^*), s^*)$ be a reciprocity*

[19] Following Rabin (1993), Falk and Fischbacher (1998) use the concept of a *psychological game* (Geanakoplos, Pearce, and Stacchetti 1989) to formulate the notion of a reciprocity equilibrium. Our formulation accomplishes the same end without requiring a notion of "subjective beliefs," which lack explanatory value in an evolutionary model.

equilibrium. Then the respondent surely accepts (i.e., $p^*(s^*) = 1$), and the proposer chooses

$$s^* = \max\left[\frac{1 + 3\rho_2 - \sqrt{1 + 6\rho_2 + \rho_2^2}}{4\rho_2}, \frac{1}{2}\left(1 - \frac{1}{\rho_1}\right)\right].$$

The theorem also holds when either or both of ρ_1, ρ_2 is zero, and if both are zero, we have the *Homo economicus* equilibrium. Note that the second expression for the equilibrium offer s^* will hold when the proposer is highly motivated by fairness, while the first expression holds if the proposer is motivated to make an offer sufficiently high so as not to be rejected.

To prove Theorem 11.4, suppose that in equilibrium the proposer offers s^* and the respondent accepts with probability p^*. Then the fairness term for the respondent is $f_2 = p^*(s^* - (1 - s^*)) = p^*(2s^* - 1)$. If the respondent accepts, then the kindness term is $k_2 = (1 - s^*) - p^*(1 - s^*)$, so the total utility from accepting is $s^* + \rho_2 p^*(2s^* - 1)(1 - p^*)(1 - s^*)$. If the respondent rejects, then $k_2 = 0 - p^*(1 - s^*)$, so the total utility is $0 - \rho_2 p^*(2s^* - 1)p^*(1 - s^*)$. The net gain from accepting is thus $\Delta_a = s^* + \rho_2 p^*(2s^* - 1)(1 - s^*)$, which is positive if $s^* \geq 1/2$. Suppose $s^* < 1/2$. Then if $\Delta_a > 0$ for $p^* = 1$, the responder will still choose $p = 1$. Otherwise let $\hat{p} = s^*/\rho_2(1 - 2s^*)(1 - s^*)$, which equates the payoffs to accepting and rejecting the offer. If $p^* < \hat{p}$, then the payoff to accepting exceeds the payoff to rejecting, so $p^* = 1$, which is a contradiction. Similarly, if $p^* > \hat{p}$, then rejecting dominates accepting, so $p^* = 0$, which is a contradiction. Thus, $p^* = \hat{p}$.

To determine s^*, let $p(s)$ be the responder's probability, derived above. Then the proposer's fairness term is $f_1 = p(s^*)(1 - s^* - s^*)$, and his kindness term is $k_1 = p(s)s - p(s^*)s^*$, so his payoff to proposing s is

$$u_1(s) = p(s)(1 - s) + \rho_1 p(s^*)(1 - 2s^*)(p(s)s - p(s^*)s^*).$$

Clearly, $u_1(s)$ is decreasing for $s > 1/2$, $s^* \leq 1/2$. Note that the smallest s such that $p(s) = 1$ satisfies $s = \rho_2(1 - 2s)(1 - s)$, which is given by $\hat{s} = \left(1 + 3\rho_2 + \sqrt{1 + 6\rho_2 + \rho_2^2}\right)/4\rho_2$. Moreover, it is easy to see that both $p(s)(1 - s)$ and $p(s)s$ are increasing so long as $p(s) < 1$, so $\hat{s} \leq s^* \leq 1/2$. This means that $p^*(s^*) = 1$. The derivative of $u_1(s)$ is $\rho_1(1 - 2s^*) - 1$, so if this is negative, then we must have $s^* = \hat{s}$. But we cannot have

$\rho_1(1-2s^*)-1>0$, or $s^*=1$, which is a contradiction. Therefore, the alternative is $\rho_1(1-2s^*)-1=0$, which means $s^*=(1-1/\rho_1)/2$, completing the proof. Q.E.D.

This theorem assumes the proposer knows the respondent's ρ_2, which accounts for the fact that offers are never refused. It is not difficult to see how to modify this by assuming the proposer knows only the probability distribution over respondent types.

As another example of a reciprocity equilibrium, let the game $\mathcal{G}$ be the Prisoner's Dilemma (§2.6, §6.11), with cooperative payoffs (b,b), mutual defect payoffs (c,c), and where a cooperator receives 0 against a defector and a defector receives a against a cooperator. We assume $a>b>c>0$. Suppose $\mathcal{G}$ is *sequential*, with One going first and choosing "cooperate" with probability p, then Two choosing "cooperate" with probability q if One cooperated, and choosing "cooperate" with probability r if One defected. We have this theorem.

THEOREM 11.5 *Suppose players in the sequential Prisoner's Dilemma game $\mathcal{G}$ have utility functions given by (11.4), where $\rho_1,\rho_2>0$ are known by both players. Then there is a unique reciprocity equilibrium (p^*,q^*,r^*) with the following characteristics.*

a. $r^*=0$.

b. *q^* is the larger of zero and $q^*=1-(a-b)/(\rho_2 ab)$.*

c. *Let $\hat{p}=(q^*b-c)/(\rho_1 a(1-q^*)(q^*b+(1-q^*)a-c))$. Then $p^*=\hat{p}$ provided this quantity is between zero and one. If $\hat{p}<0$, then $p^*=0$, and if $\hat{p}>1$, then $p^*=1$.*

Part (a) says that if One defects, Two defects as well. Part (b) says that if One cooperates and if the strength of Two's reciprocity motive ρ_2 is sufficiently strong, Two cooperates with positive probability. Also, this probability is increasing in the strength of Two's reciprocity motive, but it never reaches 100%. Part (c) is a little more complicated. The numerator is the expected gain from cooperation q^*b over defection c. If this is positive, the denominator is as well, so a selfish One (low ρ_1) will cooperate with certainty, whereas a reciprocator (high ρ_1) may not, because he is averse to giving Two a high payoff from defecting. The denominator is necessarily positive, so if the numerator is negative, no proposer will cooperate.

To prove Theorem 11.5, suppose One defected. Two's fairness term is then $f_2=r^*(-a)+(1-r^*)(c-c)=-ar^*$. His kindness term from cooperating is $k_2=a-\pi_1^*$, where π_1^* is One's equilibrium payoff given that One defected. Two's kindness term from defecting is $k_2 = c - \pi_1^*$. Two's payoff from cooperating is thus $u_2 = 0 - \rho_2 ar^*(a - \pi_2^*)$, where π_2^* is the equilibrium payoff given that One defected, and Two's payoff from defecting is $u_2 = c - \rho_2 ar^*(c - \pi_2^*)$, so clearly Two will defect with probability 1, giving $r^* = 0$. This proves (a).

Now suppose One cooperated. Two's fairness term is then $f_2 = q^*(b - b) + (1 - q^*)a = a(1 - q^*)$. His kindness term from cooperating is $k_2 = b - \pi_1^*$, where π_1^* is One's equilibrium payoff. His kindness term from defecting is $k_2 = 0 - \pi_1^*$. Two's payoff from cooperating is thus $u_2 = b + \rho_2 a(1 - q^*)(b - \pi_1^*)$. Two's payoff from defecting is $u_2 = a - \rho_2(1 - q^*)\pi_1^*$. Let $\hat{q}$ be the value of q^* that equates the two, so $\hat{q} = 1 - (a - b)/\rho_2 ab$. Clearly, $q^* < \hat{q}$ is impossible since in this case Two always cooperates, so $q^* = 1$, which is a contradiction. If $q^* > \hat{q}$, Two always defects, so $q^* = 0$, which is impossible unless $\hat{q} < 0$. We conclude that if $\hat{q} \geq 0$, then $q^* = \hat{q}$, and if $\hat{q} < 0$ then $q^* = 0$. This proves (b).

One's fairness term is

$$f_1 = p^*q^*b + (1 - p^*)q_1 - (p^*q_2 + (1 - p^*)(1 - r^*)c),$$

where $q_1 = r^*a + (1 - r^*)c$ and $q_2 = q^*b + (1 - q^*)a$. Since $r^* = 0$, this reduces to $f_1 = -p^*(1 - q^*)a$. If One cooperates, his kindness term is $k_1 = q^*b + (1 - q^*)a - \pi_2^*$, and if One defects, his kindness term is $k_1 = (1 - r^*)c - \pi_2^* = c - \pi_2^*$. Therefore, One's payoff from cooperating is $q^*b - \rho_1 p^*(1 - q^*)a(q^*b + (1 - q^*)a - \pi_2^*)$, and One's payoff from defecting is $r^*a + (1 - r^*)c - \rho_1 p^*(1 - q^*)a(c - \pi_2^*) = c - \rho_1 p^*(1 - q^*)a(c - \pi_2^*)$. Let $\hat{p}$ be the value that equates these two, so $\hat{p} = (q^*b - c)/\rho_1 a(1 - q^*)(q^*b + (1 - q^*)a - c)$. Now, if $p^* < \hat{p}$, then $p^* = 1$ since cooperating dominates defecting, and this requires $\hat{p} > 1$. If $p^* > \hat{p}$ then $p^* = 0$, so $\hat{p} < 0$. It follows that if $0 < \hat{p} < 1$, then $p^* = \hat{p}$, which completes the proof of the theorem. Q.E.D.

11.7 Altruism and Assortative Interactions

In an evolutionary model we equate welfare with fitness—the capacity to produce offspring with one's own characteristics. The altruist, by definition, becomes less fit in the process of rendering the group more fit, so evolution should entail the disappearance of the altruist. This argument has been applied to biological models where "offspring" means "children," but it applies equally to cultural models where successful behaviors are adopted by other agents. A culturally altruistic behavior is one that confers benefits on the group but is less likely to be copied by other group members than the nonaltruistic behavior.

This argument does not completely rule out altruism. Suppose there are many groups, and the altruists so enhance the fitness of the groups they are in, compared to the groups without altruists, that the former outcompete the latter, so that the average fitness of the altruist is higher than that of the selfish agent. Altruism can then proliferate. For instance, platoons with brave soldiers may survive in a situation where platoons with selfish soldiers are defeated, *despite* the fact that the brave soldiers have higher mortality than the selfish soldiers within the "brave" platoons. Overall, then, the frequency of brave soldiers in the army may increase.

To formalize this idea, suppose there are groups $i = 1, \ldots, n$, and let f_i be the fraction of the population in group i. Let π_i be the mean fitness of group i, so $\overline{\pi} = \sum_i f_i \pi_i$ is the mean fitness of the whole population. We assume groups grow from one period to the next in proportion to their relative fitness, so if f_i' is the fraction of the population in group i in the next period, then

$$f_i' = f_i \frac{\pi_i}{\overline{\pi}}.$$

Now, suppose there is a trait with frequency x_i in group i, so the frequency of the trait in the whole population is $\overline{x} = \sum_i f_i x_i$. If π_i' and x_i' are the fitness of group i and the frequency of the trait in group i in the next period, then $\overline{x}' = \sum_i f_i' x_i'$, and writing $\Delta x_i = x_i' - x_i$, we have

$$\begin{aligned}\overline{x}' - \overline{x} &= \sum f_i' x_i' - \sum f_i x_i = \sum f_i \frac{\pi_i}{\overline{\pi}} x_i' - \sum f_i x_i \\ &= \sum f_i \frac{\pi_i}{\overline{\pi}} (x_i + \Delta x_i) - \sum f_i x_i \\ &= \sum f_i \left(\frac{\pi_i}{\overline{\pi}} - 1 \right) x_i + \sum f_i \frac{\pi_i}{\overline{\pi}} \Delta x_i .\end{aligned}$$

Now, writing $\Delta \overline{x} = \overline{x}' - \overline{x}$, we can rewrite this as

$$\overline{\pi} \Delta \overline{x} = \sum f_i (\pi_i - \overline{\pi}) x_i + \sum f_i \pi_i \Delta x_i . \tag{11.5}$$

The second term is just $\mathbf{E}[\pi \, \Delta x]$, the expected value of $\pi \, \Delta x$, over all groups, weighted by the relative size of the groups. If the trait in question renders individuals bearing it less fit than other group members, this term will be negative, since $\Delta x_i < 0$ within each group. To interpret the first term, note that the *covariance* between the variables π and x is given by

$$\mathrm{cov}[\pi, x] = \sum f_i(\pi_i - \overline{\pi})(x_i - \overline{x}),$$

and since $\sum f_i(\pi_i - \overline{\pi})\overline{x} = 0$, we can write (11.5) as

$$\overline{\pi}\Delta x = \mathrm{cov}[\pi, x] + \mathbf{E}[\pi, \Delta x]. \tag{11.6}$$

This is a very famous relationship in biology, called *Price's equation*(Price 1970). Note that even if the expectation term $\mathbf{E}[\pi, \Delta x]$ is negative, so individuals with the trait are disfavored within groups, the overall effect on the growth Δx of the trait in the population can be positive, provided that the covariance term $\mathrm{cov}[\pi, x]$ is positive and sufficiently large. But the covariance will be positive precisely when *groups with high average values of the trait also have above-average fitness*—i.e., when the trait improves the fitness of the group in which it is expressed at high levels.

For an example of Price's equation, consider an evolutionary game with the stage game $\mathcal{G}$ depicted here. There are two types of agents, "selfish" and "altruist." By cooperating, an agent produces a payoff (fitness increment) $b > 0$ for his partner, at personal cost $c < b$, and by defecting an agent produces zero at zero cost. This is thus a Prisoner's Dilemma in which defection is a dominant strategy. Suppose the players pair off in each period, and each type is likely to meet its own type with probability $r \geq 0$, and a random member of the population with probability $1 - r$. If $r > 0$ we say there is *assortative interaction*. We then have *Hamilton's Law* (Hamilton 1963).

	C	*D*
C	$b-c$,$b-c$	$-c$,b
D	b,$-c$	0,0

Theorem 11.6 *Consider the evolutionary game with stage game $\mathcal{G}$, in which the degree of assortative interaction is $r \geq 0$. A small number of cooperators can invade a population of selfish actors if and only if $br \geq c$.*

To prove this theorem,[20] note that there are three types of groups, aa, as, and ss, all of size 2 at the beginning of the period. Since Price's equation

[20] There is a proof using the replicator dynamic equations, but that would not illustrate Price's equation.

remains the same if we aggregate all groups with the same internal composition, we can assume there are three groups whose fraction of total membership are f_{aa}, f_{as}, and f_{ss}, with mean fitness $\pi_{aa} = b - c$, $\pi_{as} = (b-c)/2$, and $\pi_{ss} = 0$, respectively.

To determine f_{aa}, f_{as}, and f_{ss}, let f be the fraction of altruists in the population. If r is the level of assortative interaction, the frequency of aa pairs is $f_{aa} = f(r + (1-r)f)$ and the frequency of ss pairs is $f_{ss} = (1-f)(r + (1-r)(1-f))$, so the frequency of as pairs must be $f_{as} = 1 - f_{aa} - f_{ss} = 2f(1-f)(1-r)$. Note that $f_{aa} + f_{as}/2 = f$ and $f_{ss} + f_{as}/2 = 1 - f$, as expected.

Let the trait measured by x be the "frequency of altruism." Then $x_{aa} = 1$, since in aa groups all members are altruists, and similarly $x_{as} = 1/2$, $x_{ss} = 0$. Since the fraction of altruists in aa remains 1 at the end of the period, we have $\Delta x_{aa} = 0$, and for the same reason $\Delta x_{ss} = 0$. However, the expected number of altruists in as groups at the end of the period is

$$\text{fraction of altruists} \times \frac{\text{fitness of altruists}}{\text{mean fitness}} = \frac{1}{2}\frac{-c}{(b-c)/2}.$$

Thus, $\Delta x_{as} = -c/(b-c) - 1/2 = -(b+c)/2(b-c)$. The expectation term in Price's equation then becomes

$$\begin{aligned}\mathbf{E}[\pi, \Delta x] &= f_{aa}\pi_{aa}\Delta x_{aa} + f_{as}\pi_{as}\Delta x_{as} + f_{ss}\pi_{ss}\Delta x_{ss}\\ &= f(r + (1-r)f)(b-c) \times 0\\ &\quad + 2f(1-f)(1-r)\frac{b-c}{2}\left(-\frac{(b+c)}{2(b-c)}\right)\\ &\quad + (1-f)(r + (1-r)(1-f)) \times 0 \times 0\\ &= -\frac{b+c}{2}f(1-f)(1-r).\end{aligned}$$

This is negative, as expected, becomes more negative with increasing cost c of altruism, and less negative when the degree r of assortative interaction increases. The expectation term also becomes more negative when the benefit b conferred increases, because the altruist becomes less fit by comparison with the selfish agent.

Note that

$$\overline{x} = f_{aa}x_{aa} + f_{as}x_{as} + f_{ss}x_{ss}$$

$$
\begin{aligned}
&= f(r + (1-r)f) \times 1 + 2f(1-f)(1-r) \times 1/2 \\
&\qquad + (1-f)(r + (1-r)(1-f)) \times 0 \\
&= f(r + (1-r)f) + f(1-f)(1-r) = f,
\end{aligned}
$$

so the covariance term in Price's equation becomes in this example

$$
\begin{aligned}
\mathrm{cov}[\pi, x] &= f_{aa}(\pi_{aa} - \overline{\pi})(x_{aa} - \overline{x}) + f_{as}(\pi_{as} - \overline{\pi})(x_{as} - \overline{x}) \\
&\qquad + f_{ss}(\pi_{ss} - \overline{\pi})(x_{ss} - \overline{x}) \\
&= f_{aa}\pi_{aa}(x_{aa} - \overline{x}) + f_{as}\pi_{as}(x_{as} - \overline{x}) + f_{ss}\pi_{ss}(x_{ss} - \overline{x}) \\
&= f(r + (1-r)f)(b-c)(1-f) \\
&\qquad + 2f(1-f)(1-r)\frac{b-c}{2}(1/2 - f) \\
&\qquad + (1-f)(r + (1-r)(1-f)) \times 0 \times (0 - f) \\
&= (1+r)(1-f)f(b-c)/2.
\end{aligned}
$$

Note that the covariance term increases when the level r of assortative interaction increases and when the social value $b - c$ of altruism increases. The condition for the altruistic trait to increase is $\mathrm{cov}[\pi, x] + \mathbf{E}[\pi, \Delta x] > 0$ which reduces to

$$
r \geq r_* = \frac{c}{b}. \tag{11.7}
$$

It follows that *a small number of altruists can invade a population of selfish agents, provided $r \geq r_*$*, so a positive level of assortative interaction is necessary for altruism to invade. This proves the theorem. Q.E.D.

Hamilton applied this model to altruism among kin by treating r as the biological degree of relatedness between two individuals, which can be defined as follows. Suppose individual A inherited a rare one-of-a-kind mutant from an ancestor. The *degree of relatedness* of A and another individual B is the probability that B has the same rare mutant gene.[21] For instance, since humans inherit half their genes from their fathers and half from their mothers[22] a father and a son have relatedness $r = 0.5$. This is because a rare mutant inherited by the son is equally likely to have come from his

[21] The usual definition is that the degree of relatedness is the number of genes two individuals share by common descent (i.e., inherited from the same ancestor). The two definitions are the same, when properly interpreted.

[22] Actually, this is true except for a small number of sex-related genes, which we will ignore for simplicity. We also assume that mothers and fathers are not related (i.e., there is no "inbreeding").

father or his mother. Similarly, two siblings have relatedness $r = 0.5$, since a rare gene possessed by one came from one parent, who transmitted it to the other sibling with probability 1/2. You can check that grandparents and grandchildren are related $r = 0.25$, and so on.

Suppose, then, that two individuals have a degree of biological relatedness r, and one has an altruistic mutation that leads it to cooperate, increasing its partner's fitness by b at a fitness cost c to himself. Since the partner has the same mutant gene with probability r, the expected change in the frequency of the altruistic gene is $rb - c$, so the altruistic gene increases precisely when $rb > c$.

The important point of this analysis is that *the same mechanism that accounts for altruism among kin can account for altruism among unrelated individuals, if we replace the biological process of genetic inheritance by some social process that maintains the frequency* r *of assortative interaction at a sufficiently high level.*

11.8 The Evolution of Strong Reciprocity

The stunning success of *Homo sapiens* is based on the ability of its members to form societies consisting of large numbers of biologically unrelated cooperating individuals. Neoclassical economic theory explains such cooperation using models of exchange with complete contracting. We have argued that such models do not yet adequately depict current economies (§3.17), much less those more rudimentary economies that accompanied the evolutionary emergence of our species, when the biological basis for our behavior and preferences were formed.

Game theory gives us more plausible models of social cooperation with incomplete or nonexistent contracting, in which *Homo economicus* agents build reputations by cooperating and punishing noncooperators, and in repeated interactions use threats of punishment, such as trigger strategies, to induce cooperation (see chapter 6). Reciprocal altruism of this type probably accounts for a good deal of human cooperation but is in fact quite rare among other species. How do we explain this fact? Intelligence alone is not the answer. Reciprocal altruism does require a high level of cognitive development, but since this behavior occurs in some species of vampire bats, its absence in species with at least this level of cognitive development, which includes a large number of birds and mammals, remains to be explained.

A critical weakness of reciprocal altruism is that when a social group is threatened with extinction or dispersal, say through war, pestilence, or famine, cooperation is most needed for survival. But the discount rate, which depends inversely on the probability of future interactions, increases sharply when the group is threatened. Thus, *precisely when society is most in need of prosocial behavior, cooperation based on reciprocal altruism will collapse*, for the simple reason that the discount rate will rise to levels where cooperation is no longer a Nash equilibrium. The rarity of reciprocal altruism can then be explained by the fact that *for most members of most species enough of the time, the discount rate is very high* because individual death rates are high, rates of group extinction and dispersal are high, and migration across groups occurs at a high rate.

But might not the same be said of human society during most of its evolutionary history? Since primates have not developed more than rudimentary levels of reciprocal altruism despite extremely high levels of cognitive ability, such is likely to have been the case. Perhaps the development of strong reciprocity, which leads agents to cooperate and punish noncooperators *independent of the future benefits and costs of such action*, took place precisely as a solution to the problem of high discount rates. Here is a suggestion as to how this might have occurred.

Homo reciprocans is an altruist in the sense that he improves the welfare of a group of unrelated individuals at the expense of his personal welfare. For unlike *Homo economicus*, who cooperates and punishes only if it is in his long term interest to do so, *Homo reciprocans* behaves prosocially even at personal cost. If *Homo reciprocans* is an evolutionary adaptation, it must be a considerable benefit to a group to have strong reciprocators, and the group benefits must outweigh the individual's costs in the sense of Price's equation (11.6); i.e., we must have $\mathrm{cov}[\pi, x] > -\mathbf{E}[\pi, \Delta x]$, where x is the frequency of the strong reciprocity trait and π is group fitness.

Consider an n-player public goods game (§11.4.2) in which each player has an amount c that may be kept or contributed to the "common pool." If the money is contributed, an amount $b > c$ is distributed equally among the members of the group. Thus, if k players contribute, each contributing player receives kb/n, and each noncontributing member receives $c + kb/n$. If $b/n < c$, the only Nash equilibrium is universal defection, in which each player keeps c. The Folk Theorem (§6.4) states that if this game is repeated indefinitely, full cooperation becomes a subgame perfect Nash equilibrium, provided the discount rate is sufficiently low.

We model early human society as a collection of small communities, each of which is engaged in this public goods game. Defecting is always detected and is common knowledge. When the discount factor is high enough to induce cooperation, defectors are excluded from participation in the community for a number of periods just sufficient to make defecting a suboptimal strategy, at zero cost to the community.

We suppose that in each "good" period the community will persist into the next period with probability δ^*, so δ^* is the discount factor (§6.2). In each "bad" period there is a high probability $1-\delta_*$ that the community will disband, so the discount factor is $\delta_* < \delta^*$. We suppose that the "bad" state occurs with small probability $p > 0$, and for simplicity, we suppose that the threat to the community does not affect the cost c or the return b. Suppose at the beginning of each period, prior to agents deciding whether or not to cooperate, the state of the community for that period is revealed to the members. Let π^* be the present value (total fitness) of a member if all members cooperate forever and the state of the community is "good," and let π_* be the present value of universal cooperation if the state is "bad." Then the present value before the state is revealed is $\pi = p\pi_* + (1-p)\pi^*$, and we have the following recursion equations:

$$\begin{aligned}\pi^* &= b - c + \delta^*\pi \\ \pi_* &= b - c + \delta_*\pi,\end{aligned}$$

which we can solve, giving

$$\pi^* = \frac{1 + p(\delta^* - \delta_*)}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c) \tag{11.8}$$

$$\pi_* = \frac{1 - (1-p)(\delta^* - \delta_*)}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c) \tag{11.9}$$

$$\pi = \frac{1}{1 - \delta^* + p(\delta^* - \delta_*)}(b - c). \tag{11.10}$$

Note that $\pi^* - \pi_* = \pi(\delta^* - \delta_*)$, which is strictly positive, as expected. These equations assume that the fitness of a member of a community that disbands is zero, which is thus the benchmark for all fitness values, and to which we must add an exogenous "baseline fitness" to account for the change in population of the set of communities.

When can cooperation be sustained? Clearly, if it is worthwhile for an agent to cooperate in a bad period, it is worthwhile to cooperate in a good period, so we need only check the bad-period case. The current cost of cooperating is $c-b/n$, which we approximate by c for notational convenience (the approximation is good for a large community), so the condition for cooperation is $c < \delta_*\pi$. There is a Nash equilibrium in which members thus cooperate in the good state but not in the bad when the following inequalities hold:

$$\delta^*\pi > c > \delta_*\pi, \tag{11.11}$$

which will be the case if δ^* is near unity and δ_* is near zero. We assume these inequalities hold.

Suppose community i has a fraction f_i of strong reciprocators who cooperate and punish defectors independent of the state of the community. Suppose each cooperator inflicts a total amount of harm $l_r < 1$ on defectors, at a cost $c_r < 1$ to themselves. Because of (11.11), in a bad state selfish agents always defect unless punished by strong reciprocators. If there are n_i community members, in a bad state $n_i(1-f_i)$ defect, and the total harm inflicted on those caught is $n_i f_i l_r$, so the harm per defector imposed by strong reciprocators is $f_i l_r/(1-f_i)$. The gain from defecting in (11.11) now becomes $c - f_i l_r/(1-f_i)$. Thus, if the fraction f_i of strong reciprocators is at least

$$f_* = \frac{c-\pi\delta_*}{c-\pi\delta_*+l_r}, \tag{11.12}$$

complete cooperation will hold. Note that f_* is strictly between zero and one, since the numerator, which is the gain from defecting prior to being punished by reciprocators, is positive. Also, the larger l_r, the smaller the minimum fraction f_* of reciprocators needed to induce cooperation.

If $f_i < f_*$, there will be no cooperation in a bad period (we continue to assume the parameters of the model are such that there is always cooperation in the good period). In this situation the community disbands and each member takes the fallback fitness 0. The fitness π_s of members of such "selfish" communities then satisfies the recursion equation $\pi_s = (1-p)(b-c+\delta^*\pi_s)$, which becomes

$$\pi_s = \frac{(1-p)}{1-(1-p)\delta^*}(b-c). \tag{11.13}$$

Our assumption that there is always cooperation in the good state requires that $\delta^*\pi_s > c$, which becomes

$$\frac{\delta^*(1-p)}{1-(1-p)\delta^*}(b-c) > c,$$

which we will assume holds. Note that the relative fitness benefit from being in a cooperative community is

$$d\pi = \pi - \pi_s = p\pi \frac{1-(1-p)(\delta^* - \delta_*)}{1-(1-p)\delta^*} > 0. \tag{11.14}$$

We suppose that the fraction of strong reciprocators in a community is common knowledge, and strong reciprocators punish defectors only in communities where $f_i \geq f^*$, and in doing so they each incur the fixed fitness cost c_r. We shall interpret c_r as a surveillance cost, and since punishment is unnecessary except in “bad” periods, strong reciprocators will incur this cost only with probability p, so the expected fitness cost of being a strong reciprocator is pc_r.

We will use Price's equation to chart the dynamics of strong reciprocity, which in this case says the change Δf in the fraction of strong reciprocators in the population is given by

$$\Delta f = \frac{1}{\overline{\pi}}\mathrm{cov}[\pi, x] + \frac{1}{\overline{\pi}}\mathbf{E}[\pi, \Delta x], \tag{11.15}$$

where $\overline{\pi}$ is the mean fitness of the population. Let q_f be the fraction of the population in cooperative communities, so

$$q_f = \sum_{f_i \geq f_*} q_i,$$

where q_i is the fraction of the population in community i. The fitness of each member of a community with $f_i \geq f_*$ (resp. $f_i < f_*$) is π (resp. π_s), so the average fitness is $\overline{\pi} = q_f\pi + (1-q_f)\pi_s$. Note that

$$\frac{1}{\overline{\pi}}\mathbf{E}[\pi, \Delta x] = \sum_{f_i \geq f^*} q_i f_i \frac{\pi}{\overline{\pi}}(-pc_r). \tag{11.16}$$

Let $f_c = \sum_{f_i \geq f^*} q_i f_i / q_f$, which is the mean fraction of strong reciprocators in cooperative communities. Note that

$$\begin{aligned}\pi - \overline{\pi} &= (1-q_f)(\pi - \pi_s)\\ &= (1-q_f)\left[\frac{1}{1-\delta^* + p(\delta^* - \delta_*)} - \frac{1-p}{1-\delta^* + p\delta^*}\right](b-c)\\ &\approx (1-q_f)p\pi.\end{aligned}$$

This approximation will usually be very good, since $p\delta_*$ is very small compared to $1 - \delta^*(1-p)$, and it is harmless anyway, so we will assume that $\pi - \overline{\pi} = (1-q_f)p\pi$. But then $\pi/\overline{\pi} = 1/(1-p(1-q_f))$, so (11.16) becomes

$$\frac{1}{\overline{\pi}}\mathbf{E}[\pi, \Delta x] = -\frac{pc_r q_f f_c}{1-p(1-q_f)}. \tag{11.17}$$

To evaluate the covariance term, we define $f_s = \sum_{f_i<f^*} q_i f_i/(1-q_f)$, which is the mean frequency of strong reciprocators in noncooperative communities. Also, $\pi(1-p(1-q_f)) = \overline{\pi}$, so

$$\frac{\pi}{\overline{\pi}} - 1 = \frac{(1-q_f)p}{1-p(1-q_f)}.$$

Similarly, we have $\pi_s - \overline{\pi} = -q_f(\pi - \pi_s) = -q_f p\pi$, so

$$\frac{\pi_s}{\overline{\pi}} - 1 = -q_f p\frac{\pi}{\overline{\pi}} = \frac{-q_f p}{1-p(1-q_f)}.$$

Therefore, we can evaluate the covariance term as

$$\begin{aligned}\frac{1}{\overline{\pi}}\mathrm{cov}(\pi_i, f_i) &= \sum_{f_i \geq f_*} q_i\left(\frac{\pi}{\overline{\pi}} - 1\right) f_i + \sum_{f_i<f_*} q_i\left(\frac{\pi_s}{\overline{\pi}} - 1\right) f_i \\ &= q_f f_c\frac{(1-q_f)p}{1-p(1-q_f)} - (1-q_f)f_s\frac{-q_f p}{1-p(1-q_f)} \\ &= \frac{q_f(1-q_f)p}{1-p(1-q_f)}(f_c - f_s).\end{aligned}$$

Thus, the condition for the increase in strong reciprocity is

$$(1-q_f)\left(1 - \frac{f_s}{f_c}\right) - c_r > 0, \tag{11.18}$$

and equilibrium occurs when the left-hand side of the equation is zero.

From (11.18) we get the following.

THEOREM 11.7 *Under the condition stated above, the fraction of strong reciprocators in the population lies strictly between zero and one in equilibrium. Moreover, a small number of strong reciprocators can invade a population of selfish types, provided f_s/f_c is sufficiently small, i.e., provided the strong reciprocators have a sufficiently strong tendency to associate with one another.*

Suppose communities are of size n and form randomly, the overall frequency of strong reciprocators being f. Then the expected frequency of strong reciprocators in each community will be f, with variance $f(1-f)/n$. Therefore, if $f < f_*$ and if n is large (say 100), with high probability no communities will have $f_i > f_*$, and even if some such communities exist, f_s/f_c will be very close to unity. Therefore, we have the following.

THEOREM 11.8 *Without a positive level of assortative interactions, strong reciprocators cannot invade a population of selfish types.*[23]

So let us assume that there is some way that strong reciprocators can recognize one another. Without attempting to model community formation too closely, let us simply say that communities are of equal size, and that a fraction g are formed by assortative interactions, consisting of a fraction r of strong reciprocators and a fraction $1 - r$ drawn randomly from the population. If the fraction of strong reciprocators in the population is f, then the assortative groups have a fraction $f_c = r + f(1 - r)$ of strong reciprocators. To determine f_s, note that the fraction of strong reciprocators in assortative groups is gf_c, so the fraction in randomly formed groups is $f - gf_c$, and since such groups form a fraction $1 - g$ of the total, the fraction of strong reciprocators in a randomly formed group is $f_s = (f - gf_c)/(1 - g)$. Then if assortative groups are cooperative while randomly mixed groups are not, we have $g = q_f$, and (11.18) becomes

$$\frac{r(1-f)}{r+f(1-r)} - c_r > 0. \tag{11.19}$$

This inequality holds for any value of $r > 0$ when f is very small, which is thus the condition for the invadability of strong reciprocators *however small the level of assortative interaction*. The level r of assortative interaction does, however, determine the equilibrium frequency of strong reciprocators. Setting the left-hand side of (11.19) to zero and solving for the equilibrium frequency $\hat{f}$ of strong reciprocators, we get

$$\hat{f} = \frac{r(1-c_r)}{r(1-c_r)+c_r}. \tag{11.20}$$

[23] One may think that a pattern of outmigration from cooperative groups might allow strong reciprocity to increase, but extensive analysis by population biologists fails to turn up any plausible models of this type. For an important contribution and review of the literature, see Rogers 1990.

The fraction of strong reciprocators thus varies from zero when $r = 0$ to $1 - c_r$ when $r = 1$. We may summarize this argument by saying the following.

THEOREM 11.9 *Suppose there is a degree $r > 0$ of assortative interaction among strong reciprocators. Then a small number of reciprocators can invade a population of selfish types, and the equilibrium fraction of reciprocators is given by $\hat{f}$ in (11.20).*

11.9 *Homo parochius:* Modeling Insiders and Outsiders

From the point of view of classical political philosophy, personas such as *Homo reciprocans* and *Homo egualis* are anomalous, because the behaviors they support are of ambiguous ethical value. *Homo reciprocans* is a spontaneous and often unconditional cooperator (ethically positive), but is morally judgmental and vindictive (ethically negative, at least according to liberal ethics). *Homo egualis* seeks equality, but even at the expense of pulling down everyone if that hurts the well-off more than himself. *Homo parochius*, who divides the world into *insiders* and *outsiders* according to race, ethnicity, and other ascriptive attributes of individuals, is universally condemned in modern ethical systems (although heartily affirmed in the Old Testament and other religious documents) while being embraced by a good fraction of ordinary individuals without sufficient "moral training."

Everyday observation is sufficient to convince one that the ability and willingness to divide the world into "insiders" and "outsiders" is virtually universal in us, although many good souls refuse to participate in forms of parochialism that compromise individual rights—such as racial, ethnic, and religious intolerance, or stereotyping based on gender, sexual preference, and social class. But even where such forms of discriminatory behavior are severely frowned upon, *Homo parochius* emerges in force in the form of hometown favoritism. For every New Yorker who is a fan of the Chicago White Sox, for instance, there are doubtless a thousand who are fans of the New York Yankees. And as we have suggested, such "insider-outsider" behavior is extremely easy to invoke in an experimental setting (Turner 1984).[24]

[24] Unlike the other exotic behaviors described in this chapter, that of *Homo parochius* does exist in other highly social species, such as ants, termites, and bees, where nest-mates can be clearly distinguished by chemical markers. While it is doubtful that anything akin to

By *parochialism* we mean favoring members of one's own group over members of other groups at a net material cost to oneself. The following model of parochial behavior, based on a model of reciprocity developed in Sethi and Somanathan (1999), shows that parochial behavior can be the stable outcome of an evolutionary dynamic (we do not distinguish between cultural and genetic mechanisms).

Suppose a group of n fishers share a lake that has a carrying capacity of one ton of fish per fisher per season. If fisher i exerts effort x_i, his net profit, measured in tons of fish in a season is given by

$$\pi(x_i) = x_i\left(1 - \frac{1}{n}\sum_{i=1}^{n} x_i\right) - ax_i^2, \tag{11.21}$$

where $a > 0$ is a measure of the cost of effort. Note that when total community effort is small, a unit of effort yields nearly a unit of fish, but when total effort is close to n, a unit of effort yields almost no fish. This is thus an example of the "tragedy of the commons" (§3.9, §11.4.4). If the fishers are "selfish" and choose best responses, you can easily show the following.

THEOREM 11.10 *There is a unique Nash equilibrium, in which each fisher chooses effort level*

$$x_i^* = \frac{1}{1 + 2a + n^{-1}}$$

and receives payoff

$$\pi(x_i^*) = \frac{(a + n^{-1})}{(1 + 2a + n^{-1})^2}.$$

If the community members could agree to share the catch equally and could enforce a socially optimal effort level x^{o} for each fisher, they would set

$$x^{\mathrm{o}} = \frac{1}{2(1 + a)},$$

and each fisher would have payoff

$$\pi(x^{\mathrm{o}}) = \frac{1}{4(1 + a)}.$$

"race" or "ethnicity" exists in nonhuman animals, groups of primates do offer preferential treatment to their own members.

Clearly, for large n, the lake is severely overfished, unless a is so large that the x_i^* is small, i.e., unless there is no tendency for the community to press upon the capacity limits of the lake. Note that when $a = 0$, $\pi(x^*) = n/(n+2)^2$, which is close to zero for large n. More generally, $\lim_{n\to\infty} \pi(x_i^*) = a/(1+2a)^2 \approx a$.

Suppose a fraction f of fishers are *discriminators* in the sense that instead of maximizing $\pi(x)$, they attach a positive weight α to the payoff of other discriminators, and a negative weight β to the payoff of selfish types. The "selfish" fishers (a fraction $1-f$ of the total) simply maximize their personal material payoffs. We will consider only symmetric Nash equilibria in which all discriminators choose the same effort level x_r, and all selfish types choose the same effort level x_s. Actually, as an exercise you can prove that there are no other Nash equilibria. To determine x_r and x_s, we define

$$\pi(x, y, z) = z\left(1 - \frac{(nf-1)x_r + (n(1-f)-1)x_s - x - y}{n}\right) - az^2. \tag{11.22}$$

Thus, for instance, $\pi(x, y, x)$ is the net payoff of a fisher with effort x when one other fisher has effort level y, and the remaining fishers produce at the equilibrium level. Then a selfish type chooses x to maximize $\pi(x, x_r, x)$. Solving the first-order condition for x and setting $x = x_s$, we find that the selfish fisher's optimal effort is

$$x_s^* = \frac{n(1 - fx_r)}{(1+2a-f)n+1}. \tag{11.23}$$

We find x_r by maximizing the following expression with respect to y:

$$u_r(y) = \pi(x_s^*, y, y) + \alpha(fn-1)\pi(x_s^*, y, x_r) - \beta(1-f)n\pi(x_s^*, y, x_s^*), \tag{11.24}$$

where we set $x_s = x_s^*$ in (11.22). In this equation, the first term is the catch of the discriminator who is choosing y; the second term is the catch of the other discriminators, weighted by α; and the third term is the catch of selfish types, weighted negatively by β. Solving the first-order condition $u_r'(y) = 0$, we find

$$x_r^* = \frac{n(n\beta(1+2a-f)+1)}{d_2n^2 + 2d_1n + d_o}, \tag{11.25}$$

where $d_2 = 4a^2 + 2\alpha(1+f\alpha) + (1-f)n(\alpha+\beta)$, $d_o = (1-\alpha)(n+1)$, and $d_1 = (a(2-\alpha) + f\alpha)$. Note that when $f = 1$, $x_r^* = n/((1-\alpha)(n+$

$1) + 2an + 1)$, which gives the socially optimal level of $x_r^* 1/2(1+a)$ when $\alpha = 1$, as we would expect. Substituting 11.25 in the expression for x_s^*, we find

$$x_s^* = x_r^* \frac{1 + 2an + (fn - 1)\alpha}{1 + 2an + (1 - f)n\beta}. \tag{11.26}$$

We assume a replicator dynamic (§9.2), in which the fraction f of discriminators increases when its payoff is greater than that of selfish types. The payoff to discriminators is then $\rho_r(f) = \pi(x_s, x_r, x_r)$, and the payoff to selfish types is $\pi(x_s, x_r, x_s) = \rho_s(f)$. We then can calculate that the difference in payoffs to discriminators and selfish types is given by

$$\rho_r(f) - \rho_s(f) = \frac{n((1 - f)n\beta - (fn - 1)\alpha)f_1}{d_2n^2 + 2d_1n + d_o}, \tag{11.27}$$

where $f_1 = (1 + (fn - 1)\alpha + an(2 + (fn - 1)\alpha - (1 - f)n\beta))$. The equal payoff condition $\rho_s(f^*) = \rho_r(f^*)$ is then satisfied by two values:

$$f_1^* = \frac{\alpha + n\beta}{n(\alpha + \beta)} \qquad f_2^* = -\frac{1 - \alpha + an(2 - \alpha - n\beta)}{n\alpha + an^2(\alpha + \beta)}. \tag{11.28}$$

To check for stability, we form the replicator equation

$$\dot{f} = f(\rho_r(f) - \bar{\rho}(f)),$$

where $\bar{\rho}(f) = f\rho_r(f) + (1 - f)\rho_s(f)$ is the mean fitness of the population. The Jacobian of this expression is quite complicated, but Mathematica calculated it without much complaint, and evaluating this at f_1^* we find that the equilibrium is stable, while evaluating at f_2^* (which may in fact be negative, and hence behaviorally meaningless) we find f_2^* is unstable.

Can discriminators invade a monomorphic population of self-interested fishers? We find that the Jacobian at $f = 0$ is given by

$$J_{f=0} = \frac{n(\alpha + n\beta)(1 - \alpha + an(2 - \alpha - n\beta))}{(1 + n + 2an)^2(1 - \alpha + 2an)^2},$$

which is positive when

$$\beta < \frac{1 - \alpha + (2 - \alpha)an}{an^2}. \tag{11.29}$$

A similar analysis of the case where $f = 1$ shows that the Jacobian is always strictly positive, so a monomorphic community of discriminator fishers cannot be invaded by self-interested fishers. In both the $f = 0$ and $f = 1$ cases, our inference holds only for sufficiently large n, since we assume the Jacobian does not change sign when a small number of mutants invade. We thus arrive at the following theorem.

THEOREM 11.11 *There is an evolutionary equilibrium in which the frequency of discriminators is given by* f_1^* *in (11.28) and hence is strictly positive. The effort levels of all agents are equal and are the same as the Nash equilibrium with self-interested agents exhibited in Theorem 11.10. When the frequency of discriminators is greater than* f_1^**, the Nash equilibrium with discriminators is more efficient than the Nash equilibrium with self-interested agents. Conversely, when the frequency of discriminators is positive but less than* f_1^**, the Nash equilibrium with discriminators is less efficient than the Nash equilibrium with self-interested agents.*

If n *is sufficiently large, a monomorphic population of self-interested fishers can be invaded by a small number of discriminators provided they are not "too prejudiced," in the sense of (11.29).*

Finally, if n *is sufficiently large, a monomorphic population of reciprocating fishers cannot be invaded by a small number of self-interested fishers.*

How might α and β move if they are subject to a replicator dynamic in the case of the heterogeneous equilibrium? Consider a community with a fraction f of discriminators characterized by parameters α and β. Suppose one discriminator "mutates" to a slightly larger value of α. I will not present the equations here because the calculations are complicated but (by now) straightforward (thank the Lord for Mathematica!). We have the following.

THEOREM 11.12 *If* $f > f_1^*$*, a mutant with a lower* α *has a relatively higher payoff, so under a replicator dynamic* α *will fall when the fraction of discriminators is in the efficiency-enhancing region. Conversely, if* $f < f_1^*$*, a mutant with a higher* α *has a relatively higher payoff, so under a replicator dynamic* α *will rise when the fraction of discriminators is in the efficiency-reducing region. At* $f = f_1^*$*,* α*-mutants have the same payoff as other discriminators.*

The parallel question for the parameter β has the following answer.

THEOREM 11.13 *If* $f > f_1^*$*, a mutant with a higher* β *has a relatively higher payoff, so under a replicator dynamic* β *will rise when the fraction of discriminators is in the efficiency-enhancing region. Conversely, if* $f < f_1^*$*,* *under a replicator dynamic β will fall when the fraction of discriminators is in the efficiency-reducing region. At $f = f_1^*$, β-mutants have the same payoff as other discriminators.*

Note that mutations in α and β move the dynamical system toward the equilibrium at $f = f_1^*$. However, the system is only *neutrally stable* in α and β at the equilibrium, so conditions other than those discussed in the model determine their equilibrium values, and, with them, the equilibrium f in a system in which α and β are endogenous.

a. Prove Theorem 11.10.
b. Prove Theorem 11.11 under the condition $a = 0$.
c. Prove Theorem 11.12 under the condition $a = 0$.
d. Prove Theorem 11.13 under the condition $a = 0$.