Individual Strategy and Social Structure

An Evolutionary Theory of Institutions

H. Peyton Young

PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

CONTENTS

PREFACE

THIS BOOK is a slightly lengthened version of a series of lectures I gave in June 1995 at the Summer School in Economic Theory at the Institute for Advanced Studies of Hebrew University. The goal of the book is twofold. One is to suggest a reorientation of game theory in which players are not hyper-rational and knowledge is incomplete. In particular, I dispense with the notion that people fully understand the structure of the games they play, that they have a coherent model of others' behavior, that they can make rational calculations of infinite complexity, and that all of this is common knowledge. Instead, I postulate a world in which people base their decisions on limited data, use simple predictive models, and sometimes do unexplained or even foolish things. Over time, such simple adaptive learning processes can converge to quite complex equilibrium patterns of behavior. Indeed a surprising number of classical solution concepts in game theory can be recovered via this route.

The second goal of the book is to suggest how this framework can be applied to the study of social and economic institutions. Here, I use "institution" in its everyday sense: "an established law, custom, usage, practice, organization" (Shorter Oxford English Dictionary). I take the view that institutions emerge over time from the cumulative experience of many individuals. Once their interactions coalesce into a settled pattern of expectations and behaviors, an "institution" has come into being. The theory makes qualitative predictions about the evolutionary paths that such processes tend to follow and the diversity of institutional forms that they produce. The theory also tells us something about the welfare properties of these institutions. I illustrate these ideas through a variety of simple examples, including segregated neighborhood patterns, forms of economic contracts, terms of distributive bargaining, norms of cooperation, conventions of social deference, rules of the road, and so forth. These examples are illustrative and meant to suggest directions for future work; I do not pretend to give a definitive account of the history of any one institutional form.

Like all "new" approaches, the one presented here rests on ideas that have been around for a long time. Of particular importance is the work of Thomas Schelling, who to my knowledge was the first economist to show explicitly how micro decisions by many individuals evolve into recognizable patterns of macro behavior. If this book adds anything to his, it is to provide the analytical foundations for studying these kinds of models, and to widen their sphere of application. The second shoulder on which this book rests is the work of biologists Maynard Smith and Price. Like Schelling, they showed how game theory provides the crucial link between the micro behavior of individuals and the aggregate behavior of populations. As biologists, however, they were not motivated by the idea that individuals respond rationally to their environment. Instead, they maintained, poorly adapted individuals (animal or human) are weeded out by natural selection: those who play the game well breed faster than those who do not play as well. This yields an evolutionary selection process known as the *replicator dynamic*.

To study this selection dynamic, Maynard Smith and Price (1973) introduced a new equilibrium concept known as an *evolutionarily stable strategy* (ESS). An ESS is a frequency distribution of strategies in the population that cannot successfully be invaded by a small group of mutants. Any such distribution must be a Nash equilibrium of the underlying game, but not every Nash equilibrium is an ESS. The evolutionary perspective thus provides a novel twist to a classical solution concept in game theory, and suggests how it can be strengthened. For a more complete account of this approach, the reader is referred to the pioneering work of Maynard Smith (1982) and the superb surveys by Hofbauer and Sigmund (1988) and Weibull (1995).

This book takes a different approach, one that is closer in spirit to Schelling than to the biologists. First, my interest is in economic and social phenomena, not the behavior of mice and ants. This calls for a different class of adaptive dynamics. Second, the standard solution concept in this literature, ESS, is not sharp enough for my purposes. Indeed, it was dissatisfaction with this idea that led Dean Foster and me to formulate an alternative solution concept known as stochastic stability, which is the foundation of much that follows. Roughly speaking, an equilibrium is "stochastically stable" if it is robust against persistent random shocks, not just isolated shocks, as is assumed for ESS. This leads to a much sharper notion of equilibrium (and disequilibrium) selection, as I shall show in subsequent chapters.

The book is necessarily somewhat technical, but I have tried to keep it unencumbered by lengthy proofs, which have been relegated to the Appendix. A rudimentary acquaintance with game theory is presupposed; the required elements of dynamical systems theory and Markov processes are developed from scratch. The material should be easily accessible to graduate students and professional economists, and it is not beyond the reach of advanced undergraduates. The lectures were delivered in five sessions of about an hour and a half each. The book is longer, but not by much, so it can easily serve as a module in a traditional game theory course.

Chapter 1

OVERVIEW

ECONOMIC and social institutions coordinate people's behaviors in various spheres of interaction. Markets coordinate the exchange of particular kinds of goods at specific times and places. Money coordinates trade. Language facilitates communication. Norms of etiquette coordinate how we interact socially with one another. The common law defines the bounds of acceptable behavior with respect to persons and property, and tells us what to expect when we overstep these bounds. These and many other institutions are the product, at least in part, of evolutionary forces. They are shaped by the cumulative impact of many individuals interacting with one another over long periods of time. Markets often grow up at convenient meeting places, such as a crossroads or a shady place (under the buttonwood tree). Customers come to expect particular kinds of goods to be offered there, and sellers come to meet their expectations. They also come to expect certain days and hours of operation, and particular procedures governing trade, whether posted price, haggling, or auction. These features are determined to a considerable degree by the accumulation of historical precedents, that is, by the decisions of many individuals who were concerned only with making the best trade at the moment, not with the impact of their decisions on the long-run development of that market.

A similar argument applies to economic contracts. When people rent apartments, for example, they are typically presented with a standard lease; usually the only things negotiated are price and the period of occupancy. People prefer standard contracts because they are more clearly enforceable in court than contracts that are fashioned on the spot. The accumulation of precedent makes them better defined, and hence more desirable to both parties to the transaction. But how do standard contracts become standard? The answer, evidently, is through a long period of experimentation with different forms. Eventually one form becomes standard and customary for a given type of transaction (in a given locale), not necessarily because it is optimal, but because it serves the purpose reasonably well and it is what everyone has come to expect. It is now an institution that coordinates behaviors, and to deviate from it would be costly.

A similar argument can be made for a great variety of social and economic institutions—language, codes of dress, forms of money and credit, patterns of courtship and marriage, accounting standards, rules of the road. In most cases, no one willed them into being: they are what they are due to the accumulation of precedent; they emerged from experimentation and historical accident.

Of course not all institutions can be explained in this way. Some were created by edict. In the Middle Ages, market towns were often established by royal charter. Rules of the road are enshrined in statutes. Accounting standards are regulated by official or semiofficial bodies. Languages are taught from standard dictionaries and grammar books. When we look more deeply into the matter, however, we find that these codifications were often just a way of ratifying practices that had already come into being through evolution. Furthermore, codifications and edicts do not stop evolution in its tracks: market towns come and go, accounting standards change, dictionaries and law books are always being rewritten.

The notion that economic institutions and patterns of behavior can be explained as the product or outcome of many individual decisions is scarcely a new idea in economics. It is perhaps most prominently associated with members of the Austrian school, notably Menger, von Hayek, and Schumpeter, though elements of the approach are implicit in the writings of earlier authors, including Adam Smith, David Hume, and Edmund Burke.[1]

What are the features that distinguish the "evolutionary" from the classical point of view in economics? One is the status accorded to equilibrium; the other is the status accorded to rationality. In neoclassical economics, equilibrium is the reigning paradigm. Individual strategies are assumed to be optimal given expectations, and expectations are assumed to be justified given the evidence. We, too, are interested in equilibrium, but we insist that equilibrium can be understood only within a dynamic framework that explains how it comes about (if in fact it does). Neoclassical economics describes the way the world looks once the dust has settled; we are interested in how the dust goes about settling. This is not an idle issue, since the business of settling may have considerable bearing on how things look afterwards. More important, we need to recognize that the dust never really does settle—it keeps moving about, buffeted by random currents of air. This persistent buffeting by random forces turns out to be an essential ingredient in describing how things look on average over long periods of time.

The second feature that differentiates our approach from standard ones is the degree of rationality attributed to economic agents. In neoclassical economic theory—especially game theory—agents are assumed to be hyper-rational. They know the utility functions of other agents (or the probability that other agents have these utility functions), they are fully aware of the process they are embedded in, they make optimum long-run plans based on the assumption that everyone else is making optimum long-run plans, and so forth. This is a rather extravagant and implausible model of human behavior, especially in the complex, dynamic environments that economic agents typically face. Moreover it represents a peculiar aberration from traditional ways of thinking in economics. One of the central messages of the pure theory of exchange, for example, is the ability of prices and markets to coordinate economic activity *without* assuming that agents are anything more than naive optimizers acting on limited information.

In this sense, our point of view represents a return to older traditions in economics. Agents adapt—they are not devoid of rationality—but they are not hyper-rational. They look around them, they gather information, and they act fairly sensibly on the basis of their information most of the time. In short, they are recognizably human. Even in such "low-rationality" environments, one can say a good deal about the institutions (equilibria) that emerge over time. In fact, these institutions are often precisely those that are predicted by high-rationality theories—the Nash bargaining solution, subgame perfect equilibrium, Pareto-efficient coordination equilibria, the iterated elimination of strictly dominated strategies, and so forth. In brief, evolutionary forces often *substitute* for high (and implausible) degrees of individual rationality when the adaptive process has enough time to unfold.

Let us add a little more meat to this rather skeletal outline. Recall that our general objective is to show how economic and social institutions emerge from the interactive decisions of many individuals. To talk about this idea rigorously, we need a model of how individuals interact at the micro level. This is naturally provided by a *game*, which describes the strategies available to each player and the payoffs that result when they play their strategies. Obviously the form of the game depends on the interactive situation we are trying to model. To illustrate ideas, we shall usually rely on games having a relatively simple structure, such as coordination games and bargaining games, but the theory extends to all finite-strategy games, as we show in Chapter 7.

The framework differs from traditional game theory in several crucial respects, however. First, players are not fixed, but are drawn from a large population of *potential* players. Second, the probability that individuals interact depends on exogenous factors, such as where they live, and more generally on their proximity in some suitably defined social space. Third, agents are not perfectly rational and fully informed about the world in which they live. They base their decisions on fragmentary information, they have incomplete models of the process they are engaged in, and they may not be especially forward looking. Still, they are not completely irrational: they adjust their behavior based on what they think other agents are going to do, and these expectations are generated endogenously by information about what other agents have done in the past. On the basis of these expectations, the agent takes an action, which in turn becomes a precedent that influences the behavior of future agents. This creates a feedback loop of the following type:

$$\begin{array}{ccc} \text{Precedents} & \longrightarrow & \text{Expectations} \\ \nwarrow & & \swarrow \\ & \text{Actions} & \end{array}$$

Finally, we assume that the dynamic process is buffeted by random perturbations that arise from a variety of factors, such as exogenous shocks or unpredictability in people's behavior. These shocks play a role similar to that of mutations in biology by constantly testing the viability of the current regime. Moreover, they imply that *the evolutionary dynamic never settles down completely; it is always in flux.* The novel element of the approach from a technical standpoint is to show explicitly how to analyze the long-run behavior of such processes.

To illustrate these ideas concretely, consider the following variant of Schelling's model of neighborhood segregation patterns (Schelling, 1971, 1978). There are two types of people—As and Bs—who choose where they want to live. Their utility for a given location depends on the composition of the neighborhood, that is, on the mixture of As and Bs around them. The situation is depicted in Figure 1.1, where each circle represents a location. Let us suppose that an individual is *discontent* if his two immediate neighbors are unlike himself; otherwise, he is *content*. An equilibrium configuration is one in which no two individuals want to trade places. In other words, there is no pair such

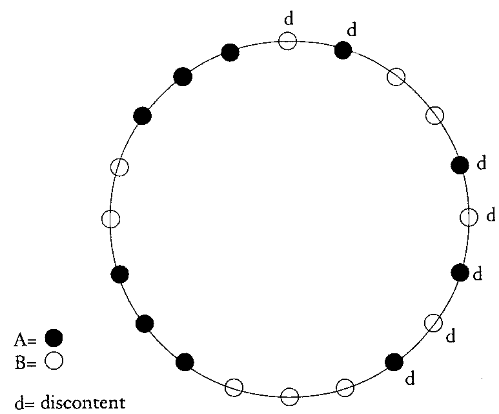

Figure 1.1. A random distribution of two types of agents on the circle.

that one (or both) is currently discontent and both would be content after they trade locations. (If only one person is discontent beforehand, we can imagine that she compensates the other to move, so that both are better off after the move than they were before.)

We claim that if there are at least two people of each type, then in equilibrium no one is discontent. To see this, suppose to the contrary that an A is surrounded by two Bs (... BAB ...). Moving clockwise around the circle, let B* be the last B-type in the string of Bs who follow this A, and let A* be the person who follows B*:

$$\ldots BAB \ldots BB^*A^* \ldots.$$

Since there are at least two agents of each type, we can be sure that A* differs from the original A. But then the original discontented A could switch with B* (who is content), and both would be content afterwards. Thus we see that the equilibrium configurations consist of those arrangements in which everyone lives next to at least one person of their own type. No one is "isolated."

In general there are many types of equilibrium configurations. Some consist of small enclaves of As and Bs scattered around the landscape; others are completely segregated in the sense that all As live on one side of the circle, and all Bs on the other side (see Figure 1.2).

So much for the equilibrium analysis. What happens when the process begins in an out-of-equilibrium situation—will equilibrium even-

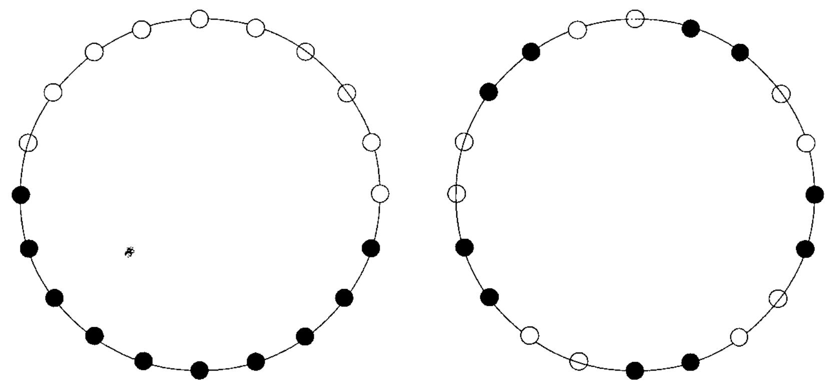

Figure 1.2. Two equilibrium configurations: one segregated, the other integrated.

tually be reached? Consider the following adjustment dynamic. Time is broken into discrete periods. In each period, two people meet at random and have a conversation about their respective neighborhoods. If they find that they can trade places to advantage, they do so. The reader may check that from any initial state, there always exists some sequence of advantageous trades that leads to an equilibrium. Since the number of states is finite and the probability of following such a path is positive, the process will eventually find its way to some equilibrium with probability one. Figure 1.3 illustrates a series of adjustments of this type. Note that the adjustment process is not fully predictable: where it ends up depends on where it starts and on the order in which people happened to trade. One can therefore speak of the probability of reaching various outcomes from some initial state, without knowing which one will in fact materialize. Such processes are sometimes said to be "path-dependent."[2] Of course, any nontrivial stochastic process is path dependent in the sense that different paths will be followed depending on the outcome of chance events. A more telling definition of path dependency is that, with positive probability, the process follows paths that have different long-run characteristics (the process is "nonergodic"). The location model described above is path dependent in this sense, because the paths end up in different equilibrium configurations.

A further complication arises from the fact that people do not always act the way our models say they do. Suppose, for example, that there is a small probability that a pair of individuals who could gain from

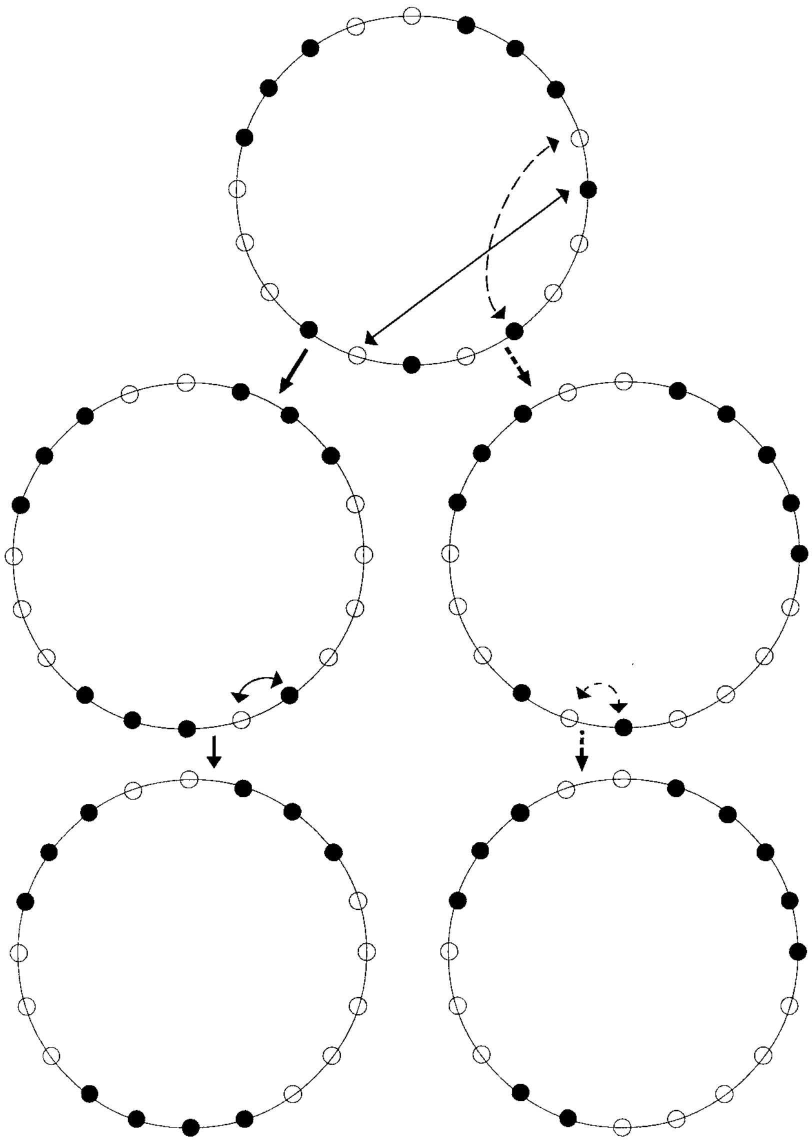

Figure 1.3. Alternative evolutionary paths from same starting point.

trading places will nevertheless fail to do so; similarly, there is a small probability that a pair who cannot gain from trading places will trade anyway. This assumption expresses the simple fact that we do not know all the reasons why people act the way they do. Our behavioral model is incomplete, and actions need to be modeled as random variables. These kinds of uncertainties have important implications for the behavior of the process. In particular, the process is *ergodic*, that is, its long-run average behavior is essentially independent of the path taken; furthermore it is independent of the initial conditions.[3] Using techniques that we shall develop in Chapter 3, we shall show that the process spends almost all of its time in one of the completely segregated states where all As live on one side of the circle and all Bs on the other side.

Note that nobody intended this outcome; it arises because people optimize locally and do not worry about the effect of their actions on the long-run properties of the system. Note also that the process keeps on evolving—it never "ends up" somewhere—because the residential pattern of As and Bs is constantly shifting: sometimes As will live in the north, sometimes in the south; segregated neighborhoods will eventually become integrated, later they will become segregated again, and so forth. This represents a fairly accurate picture of reality.

While we cannot predict the dynamic path that such a process will follow, we can estimate the *probability* with which different kinds of residential patterns will be observed over the long run. In the present example, the answer is all too familiar: when people have a preference for some neighbors who are similar, and everyone is left to their own devices, completely segregated neighborhoods are more likely to emerge than any other pattern. In fact, this remains true even if everyone *prefers* to live in a mixed neighborhood (where one neighbor is similar and the other is different), than to live in a neighborhood surrounded by their own kind. In this case the segregated states may be far from optimal, but in the long run they are the most likely states.

The analytical techniques that allow us to prove this result are developed in subsequent chapters. There is, however, an important concept at work here that can be stated without the formal apparatus: when an evolutionary process is subjected to small, persistent stochastic shocks, some states occur much more frequently than others over the long run. These states are said to be *stochastically stable* (Foster and Young, 1990). Later we shall show how to compute the stochastically stable states explicitly. What bears emphasizing here is that stochastic stability is a considerably sharper (and more general) criterion of equilibrium selection than such established notions in the literature as "evolutionary stable strategy" and "risk-dominant equilibrium," though it is related to them in special cases, as we shall see in subsequent chapters.

To illustrate the kinds of questions we can address using this concept, let us briefly consider several further stylized examples. One economic institution that clearly has an evolutionary flavor is the choice of medium of exchange.[4] History reveals the great variety of goods that societies have adopted as money: some used gold or silver; some, copper or bronze; others used beads; still others favored cattle. In the early stages of economic development, we can conceive of the choice of currency as growing out of individual decisions that gradually converge on some norm. Once enough people in a society have adopted a particular currency, everyone else wants to adopt it, too.

At the most basic level, this kind of decision problem can be modeled as a coordination game. Suppose that there are two choices of currency: gold or silver. At the beginning of a period, each person must decide which currency to carry (we assume that carrying both is too costly). During the period, each person meets various other people in the society at random, and they can trade only if they are both carrying the same currency. Thus the decision problem at the beginning of the period is to choose the currency that one thinks will be chosen by a majority of the others.

Schematically, we can model this situation as follows. Let p^t be the proportion in the population choosing gold at time t, and let $1 - p^t$ be the proportion choosing silver. In period $t + 1$, some people reconsider what they are doing (or they die and are replaced by people who must make a new decision). For the sake of simplicity, suppose that exactly one person, drawn at random from the general population, reconsiders during each period. Assume that the properties of gold and silver make them equally desirable as currencies. (We shall relax this assumption in a moment.) Then our decisionmaker chooses gold if $p^t > .5$ and chooses silver if $p^t < .5$. If $p^t = .5$, we can assume that the decisionmaker continues to do whatever he was previously doing because of inertia.[5] All of this happens with high probability, say, $1 - \varepsilon$. But with probability $\varepsilon > 0$ a person chooses gold or silver at random, that is, for reasons external to the model.

Qualitatively, this process evolves in the following manner. After an initial shakeout, the process converges quite rapidly to a situation in which most people are carrying the same currency—say, gold. This norm will very likely stay in place for a considerable period of time.

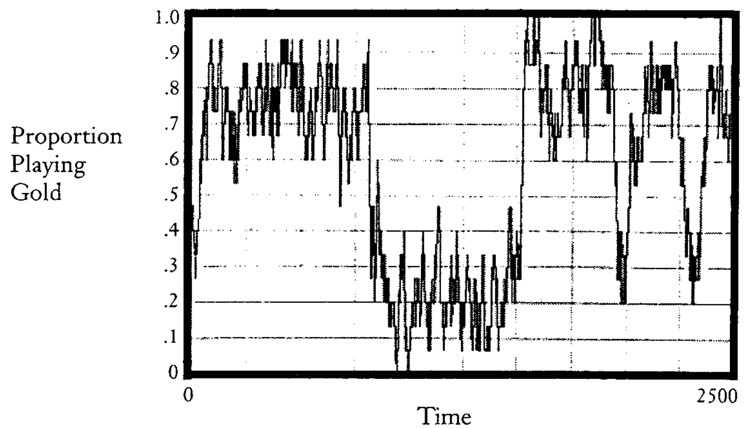

Figure 1.4. The currency game with equal payoffs, population size 10, and $\varepsilon = .5$.

Eventually, however, an accumulation of random shocks will "tip" the process into the silver norm. These tipping incidents are infrequent compared to the periods in which one or the other norm is in place (assuming ε is small). Moreover, once a tipping incident occurs, the process will tend to adjust quite rapidly to the new norm. This pattern—long periods of stasis punctuated by sudden changes of regime—is known in biology as the *punctuated equilibrium effect*. The evolutionary model we have described above predicts that a similar phenomenon characterizes shifts in economic and social norms. Figure 1.4 illustrates this phenomenon for the currency game when the two currencies have equal payoffs.

Now let us ask what happens when one currency is inherently a little better than the other. Suppose that gold is somewhat preferred because it does not tarnish as easily as silver. Then the decision problem at the individual level is to choose gold if $p^t > \gamma$, and to choose silver if $p^t < \gamma$, where γ is some fraction less than .5 but larger than 0. Now the process follows a path that looks like Figure 1.5. Over the long run, there is a bias toward gold; that is, at any given time, the society is more likely to have adopted the gold standard than to have adopted the silver standard. This is not surprising. What is perhaps surprising is that the bias becomes larger the smaller the random perturbations are. Figures 1.6 and 1.7 show characteristic sample paths for $\varepsilon = .10$ and $\varepsilon = .05$. It is clear that the smaller the value of ε, the more likely it is that the process is in a gold standard phase, and the longer are the

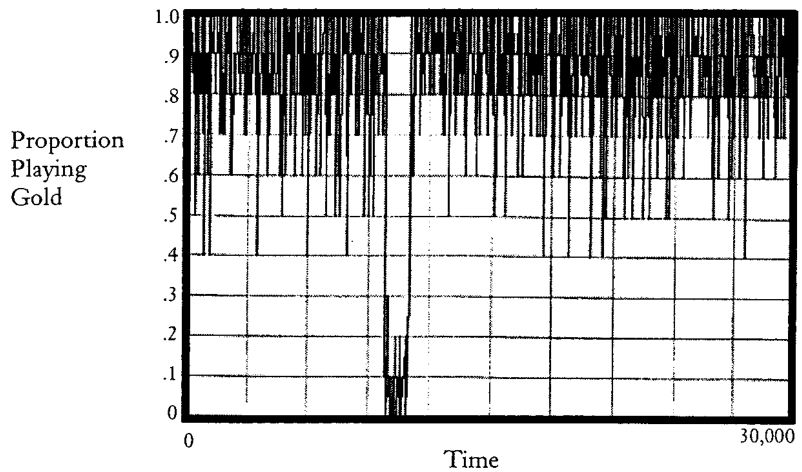

Figure 1.6. The currency game with asymmetric payoffs, population size 10, and $\varepsilon = .10$.

tial information" model has long-run properties very similar to those of the full information model, as we shall show in later chapters.) Just as in the currency game, the process oscillates between periods in which first one technology and then the other is dominant. However, if one technology is better (even a little better) than the other, the model predicts that over the long run the process will find its way to the superior technology, which will tend to stay in place for longer periods of time than will the inferior one. In our terminology, the superior technology is stochastically stable.

This result must be interpreted with some caution, however, because the long run in such a model can be very long indeed compared to the rate of technological change. By the time the process finds its way to the superior technology, the nature of the two technologies may have changed entirely. As Arthur (1989) has argued, what matters from a practical standpoint is who seizes the largest share of the market initially, for this confers an advantage on the leader that can be very difficult to overcome in the short or intermediate run. To put it another way, if an inherently inferior technology gets a head start on an inherently superior one as a result of chance events, and if there are strong networking effects, then the inferior technology is likely to hold onto its lead for a long period of time before stochastic forces displace it in favor of the superior technology.[7]

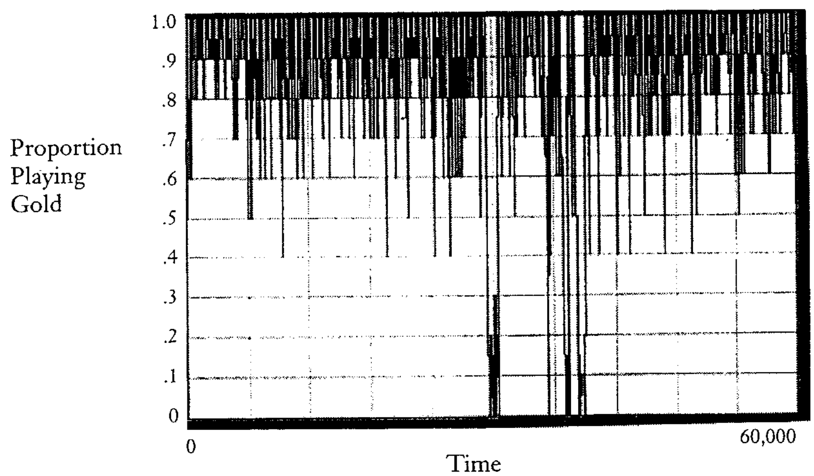

Figure 1.7. The currency game with asymmetric payoffs, population size 10, and $\varepsilon = .05$.

Thus, from a short-run perspective, a key property of the system is its *inertia*, that is, the expected waiting time until the process tips from the less favorable to the more favorable regime. While this depends in part on the comparative superiority of one technology over the other, it also depends on (i) the size of the customer base, (ii) the amount of information on which customers base their decisions, (iii) the magnitude of the stochastic perturbations, and (iv) the extent to which customers gather information locally or globally. When people have a large information base and they interact globally, the inertia can be enormous: once an inferior technology gets a jump on a superior one, it can take almost forever for evolutionary forces to dislodge it. On the other hand, when people base their decisions on relatively little information, and they interact mainly with small groups of neighbors, the process can find its way to the optimum technology relatively quickly. Note, however, that low inertia is a two-edged sword: while it reduces the waiting time to get to the optimum outcome, the optimum does not remain in place as long before it risks being displaced (by further shocks) in favor of a suboptimal outcome.

This discussion raises an important issue in evolutionary models, namely, the time scale in which events unfold. In the processes we shall study, time is measured in discrete periods that correspond to distinct events. For example, each interaction between a pair of individuals might mark the beginning of a new period. When the population is large and people interact often, thousands or even millions of such events might be compressed within a short period of real time, such as an hour or a day. It is therefore not meaningful to make statements about short- versus long-run phenomena without a metric for translating event time into real time.

An implicit assumption in these models is that some parameters are changing much more slowly than others, so that the former can be viewed as fixed compared to the latter. When we model strategic behavior in a game, for example, we usually assume that the payoff structure remains fixed while the expectations of the players change. This assumption makes more sense in some cases than in others. In the competition between computer technologies, for example, payoffs may be changing so quickly that they need to be viewed as a dynamic element of the system. In other situations, the game may evolve very slowly. Consider the problem of which side of the road to drive on. At the micro level, this can be viewed as a game between two oncoming vehicles: both drivers want to coordinate on the same convention—left side or right side—in order to avoid an accident. Whether the game is played between horse-drawn carriages or high-speed automobiles, it is essentially a coordination game whose payoff structure does not change much over time. (Of course, the absolute payoffs change—the disutility of an accident increases as vehicles become faster—but all that matters is that the two competing conventions have equal payoffs.)

Rules of the road would therefore seem to be a good vehicle for studying the long-run properties of convention formation. Moreover, the history of left-right driving conventions in Europe exhibits the qualitative patterns that an evolutionary model would predict.[8] In the early stages, when there was relatively little traffic on the roads and its range was limited, conventions grew up locally: a city or province would have one convention, while a few miles down the road another jurisdiction would have the opposite one. As use of the roads increased and people traveled further afield, these local rules tended to congeal first into regional and then into national norms, though for the most part these norms were not codified as traffic laws until well into the nineteenth century. In areas with highly fragmented jurisdictions, the congealing process took longer, as an evolutionary model would predict. Italy, for example, was characterized by highly localized left-right driving rules until well into the twentieth century.

Once conventions become established at the national level, the interactions are between countries, who are influenced by their neighbors: if enough of them follow the same convention, it pays to follow suit. Over time, we would expect a single convention to sweep across the board. While this intuition is essentially correct, it ignores the effect of idiosyncratic shocks, which can displace one convention in favor of another. Remarkably, just such a shock occurred in the history of European driving: the French Revolution. Up to that time, it was customary for carriages in France as well as in many other parts of Europe to keep to the left when passing. This meant that pedestrians often walked on the right to face the oncoming traffic. Keeping to the left was therefore associated with the privileged classes, while keeping to the right was considered more "democratic." Following the French Revolution, the convention was changed for symbolic reasons. Subsequently Napoleon adopted the new custom for his armies, and it spread to some of the countries he occupied.

From this point forward, one can see a gradual but steady shift—moving more or less from west to east—in favor of the right-hand rule. For example, Portugal, whose only border was with right-driving Spain, converted after World War I. Austria switched province by province, beginning with Vorarlberg and Tyrol in the west and ending with Vienna in the east, which held out until the Anschluss with Germany in 1938. Hungary and Czechoslovakia also converted under duress at about this time. The last continental European country to change from left to right was Sweden in 1967. Thus we see a dynamic response to an exogenous shock (the French Revolution) that played out over the course of almost two hundred years.

Of course, whether people drive on the left or the right is not particularly consequential in terms of social welfare. What matters is that society have an established convention in which expectations and behaviors are in equilibrium. It should be borne in mind, however, that some equilibria may be quite undesirable from the standpoint of social welfare. Indeed, some games are perverse in the sense that *everyone* does poorly in equilibrium. (Prisoner's dilemma is a prominent example.) Evolutionary arguments do not negate this problem; they can explain how bad equilibria come about, but they do not eliminate them.

Consider, for example, a situation in which two individuals cooperate to produce a joint product. They can either work hard or shirk. If both work, their output is high; if both shirk, their output is low. If one works and the other shirks, the output is the same as if both had shirked, but the one who worked put in more effort (in vain) and hence is less well off than the one who shirked.[9] Assuming equal division of the spoils, we have a game with the following payoff structure (the specific numbers are unimportant):

	Work	Shirk
Work	10, 10	0, 7
Shirk	7, 0	7, 7

If you expect your partner to work, it pays to work too; if you expect your partner to shirk, it is better to shirk. Thus two different norms can emerge: work or shirk. It turns out that in an evolutionary model of the type described above, shirking is stochastically stable: it is more likely to be the norm at any given time than working. The intuitive reason is that there is always some random variation in people's strategies, and hence some uncertainty about what one's partner will do. In such an environment, shirking is a safer strategy than working. Suppose, for example, that one believes that there is a greater than 30 percent chance that one's partner will shirk. Then it is best to shirk, too. On the other hand, one must believe that there is a greater than 70 percent chance that one's partner will work to make one want to work. In the terminology of Harsanyi and Selten (1988) the shirk norm is *risk dominant*.

But why would one believe that there is a 30 percent chance that one's partner is going to shirk when almost everyone is currently working? The answer has to do with variability in the population. Even if most people currently work hard, there will almost always be a few people who for some reason or other choose to shirk. Now suppose that someone outside this group of shirkers happens to interact with them for several periods. (This is particularly likely to happen if the shirkers are concentrated in some neighborhood.) This person will come to believe (given his or her limited information) that a sizable fraction of people shirk, which will induce him or her to shirk, too. This action will then be noticed by other people, which will further reinforce the idea that people are shirking. Shirking spreads by contagion. Of course, the process can also go in the reverse direction, whereby the work norm spreads by contagion. The point about the 30 percent–70 percent comparison is that it takes longer for accumulated variations to reach the 70 percent threshold than to reach the 30 percent threshold, so the waiting time to get from shirk to work is longer than the other way around.

In general, the inertia of the system—the waiting time needed to tip from one norm to another—depends in a fairly complex way on the size of the population, the extent to which people interact with their neighbors or with those far away, the amount of the information they gather, and so forth. This issue is explored in more detail in Chapters 4–6, where we show that under a wide variety of conditions, the risk-dominant norm in a 2×2 game is stochastically stable. Over the long run, it has the evolutionary advantage whether or not the outcome is socially optimal. However, while stochastic stability and risk dominance usually coincide in 2×2 interactions, they are not the same thing in general, as we show in Chapter 7.

Stochastic evolutionary models are able to make quantitative predictions about the evolution of norms and institutions (sometimes surprisingly sharp ones), but their greatest significance lies in their qualitative properties. They have a different "look" and "feel" than equilibrium models. These qualitative features are sufficiently distinctive, in fact, that there is a reasonable prospect that they could be tested against empirical data. Although this task is well beyond the scope of the present book, we can suggest several features that, in principle, are testable.

To illustrate, consider a collection of distinct societies whose members do not interact with one another. Over time, each will develop distinctive institutions to cope with various types of economic and social coordination (forms of contracts, work norms, conventions of social behavior, and the like). We can think of these institutions as particular equilibrium outcomes in a game that has many (potential) equilibria. At any given time, a given society is likely to be "near equilibrium" in the sense that almost everybody follows the behavioral pattern that is expected of them and almost everybody wants to follow this behavior given the behavior they expect of others. Note that we say *almost* instead of *all*: there will inevitably be some misfits and nonconformists who do not follow established patterns. Moreover, according to the theory, these mutant types play a crucial role in promoting long-run change: occasionally they become numerous enough to tip society from one near-equilibrium to another. In particular, different societies may be near different equilibria at any given point in time due to the vagaries of history. This fact has two general implications. On the one hand, it says that two people in similar roles are more likely to exhibit similar behaviors if they come from the same society than if they come from different societies, assuming all other explanatory variables are held constant. This is the *local conformity* effect. On the other hand, because the same process of adaptation is operating simultaneously in all of the societies, the frequency distribution of institutional forms will be fairly stable and predictable over the long run. In particular, the theory predicts that some institutions are inherently more *stable* or *durable* than others in the presence of stochastic shocks. Once established, they tend to persist for longer periods of time. Over the long run, these institutions will occur with higher frequency among the various villages. When the stochastic shocks are small, the *mode* of this frequency distribution will tend to be close to the stochastically stable institutions predicted by theory.

These two effects can in principle be identified from cross-sectional data. A third qualitative prediction, however, concerns the look of the evolutionary path and will only be revealed by time series data. The process will tend to exhibit long periods of stasis in which it is close to some equilibrium, punctuated by relatively brief periods in which the equilibrium shifts in response to stochastic shocks. We call this the *punctuated equilibrium* effect.[10] It is a well-known feature of residential segregation patterns, but also occurs in other contexts and on different time scales. The spread of right-side driving in continental Europe appears to be one example where the tipping process took two centuries to run its course. Residential patterns, by contrast, often tip in a few years.

Organization of the Book

The book's argument is organized as follows. Chapter 2 discusses various models of adaptive behavior at the individual level. These include replicator dynamics, reinforcement models from psychology, imitation, and best-reply dynamics. For the sake of expositional clarity, we choose to focus on best-reply models, although a similar analysis could be carried out for other classes of adaptive rules. This choice is also governed by the fact that best-reply models have a firm foundation in the theory of individual choice. We can therefore address a variety of issues that are beyond the scope of the other frameworks, including the effects of risk aversion and amount of information on the evolutionary selection process.

The adaptive model is an elaboration of the feedback loop described above. Players develop expectations about others' behavior based on precedent—on information about what other people have done in the past. This information is typically fragmentary and incomplete, because a given person will generally know only a small proportion of the relevant precedents, which he learns through his social network. Furthermore memory is bounded: players do not know (or perhaps do not care) about things that happened long ago; only recent events matter. On the basis of her information, a player forms a simple statistical model of how others are likely to behave. Usually she chooses a best reply given these expectations, but sometimes she makes arbitrary or unexplained choices. This simple model of adaptive behavior forms the basis of all the subsequent analysis.

In Chapter 3 we develop a conceptual framework for studying stochastic models of learning and adaptation more generally. The chapter opens with a discussion of asymptotic stability in deterministic dynamical systems, which is the standard concept in much of the evolutionary games literature. It is argued that this concept is not satisfactory as a predictor of long-run behavior when the system is subjected to persistent stochastic shocks (as almost all such systems are). The key concept of *stochastic stability* is introduced (Foster and Young, 1990). Loosely speaking, the stochastically stable states of an ergodic stochastic process are those states that occur with nonnegligible probability when the size of the stochastic perturbations is arbitrarily small. We develop a general framework for computing the stochastically stable states of any process that is Markovian, has stationary transition probabilities, and operates on a finite state space. The approach is illustrated with the neighborhood segregation model.

Chapter 4 applies this technique to the study of two-person coordination games in which each player has just two strategies. In this case, the stochastically stable outcome coincides with Harsanyi and Selten's concept of risk dominance. We explore the implications of this result for a variety of examples, including the currency game, the work-shirk game, and games of social etiquette. We then show how to compute the long-run distribution explicitly as a function of the population size and the error rate. This analysis shows that the selective bias in favor of the risk-dominant equilibrium is quite marked even when the stochastic perturbations are large, provided the population of players is also large.

In Chapter 5 we consider various refinements and embellishments of the basic learning model. First, we examine the effects of having more or less information, that is, whether having a large or a small information base is advantageous in the long run. It turns out that the answer is complex, and that having more information is an advantage in some kinds of games, but not in others. Then we show how to analyze situations where the populations are heterogeneous with respect to the individual payoff functions and the amount of information the players have. Next, we investigate the sensitivity of the selection results to different ways of modeling stochastic perturbations. In particular, we consider the case where players deviate from best reply at different rates. We also consider the possibility that they choose a nonbest reply with a probability that decreases as the prospective loss of payoff increases (Blume, 1993). Under either modification, the stochastically stable outcome in a symmetric 2×2 coordination game remains the risk-dominant equilibrium, just as in the case of uniform random errors.

Finally, we examine the situation where memory is not bounded, and players attach equal importance to all precedents no matter how dated they are. This process exhibits substantially different long-run behavior than does the process with bounded memory. In particular it is not ergodic: the process converges to an equilibrium (or near-equilibrium) regime with a probability that depends on the initial state. We argue, however, that this result is of mainly theoretical interest, since in practice past actions do *not* carry the same weight as recent ones.

Chapter 6 analyzes situations where players are located in a geographic or social space, and interact only with their "neighbors."[11] We posit a perturbed best-reply process in which each player chooses an action with a probability that decreases exponentially the lower its expected payoff is, given what his neighbors are doing. This formulation allows us to represent the long-run probability of different states as a Gibbs distribution. Once again, we find that the stochastically stable outcome in a symmetric 2×2 game occurs when everyone plays the risk-dominant equilibrium. This framework is also convenient for studying the *inertia* of the system—how long it takes (in expectation) for the process to reach the stochastically stable outcome from an arbitrary initial state. Extending previous work of Ellison (1993), we show that the inertia of a local interaction system can be dramatically lower than the inertia of a system in which everyone interacts with everyone else. In particular, if everyone interacts with a sufficiently small, close-knit group, the inertia of the process is bounded above independently of the total population size.

In Chapter 7, we extend the analysis to arbitrary finite n-person games. Here the key concept for studying the long-run selection process is sets of strategies that are closed under best replies. Specifically, a *curb set* is a Cartesian product set of strategies that includes all best replies to all possible probability mixtures over strategies in the set (Basu and Weibull, 1991). A *minimal curb set* is one that contains no smaller curb set. Building on work of Hurkens (1995), we show that in a generic n-person game, adaptive learning selects a minimal curb set; that is, in almost all n-person games there is a unique minimal curb set whose strategies are stochastically stable. We describe the general method for computing this set. In some classes of games the minimal curb sets correspond one to one with the strict Nash equilibria, in which case the learning model yields a theory of equilibrium selection. Moreover, there are important cases where the equilibria selected in these low-rationality environments are the same as those that emerge in the high-rationality world of classical theory.

Chapter 8 considers one such example, the Nash bargaining solution. In the traditional noncooperative model of bargaining, two players take turns making offers about how to divide a pie between themselves (Stahl, 1972; Rubinstein, 1982). If a player refuses an offer, the other player makes a counteroffer. If a player accepts an offer, the game ends. Assuming there is a small probability that the bargaining will "break down" after each refusal, the subgame perfect equilibrium outcome of this game is close to the division that maximizes the product of the players' utilities (the Nash bargaining solution). Note that this argument relies on the assumption that both players understand the structure of the game, that their utility functions are common knowledge, and that their rationality is common knowledge.

The evolutionary model dispenses with all three of these assumptions. As above, we assume that agents in the current period have expectations that are shaped by the precedents they know about from earlier periods. Each player normally demands an amount that has the highest expected payoff given his belief about how much the other will demand, though sometimes the demands are idiosyncratic. If the bargainers within each population have the same degree of risk aversion and the same amount of information, the stochastically stable outcome is close to the Nash bargaining solution. We therefore obtain a classical solution concept without extravagant assumptions about the agents' level of rationality or the degree of common knowledge. Moreover, when agents are heterogeneous in their characteristics, the model yields a novel generalization of the Nash solution.

A distributive bargain may be viewed as a primitive contract between two parties about dividing a pie. Chapter 9 extends the analysis to the evolution of contracts more generally. A *contract* expresses the terms that govern the relations between people. Its terms may be quite explicit, as in a bank loan, or almost entirely implicit, as in a marriage. Whether implicit or explicit, there is a tendency for contracts to follow standard formats. It is a commonplace observation, however, that what is standard in one society may not be standard in another. This fact suggests that evolutionary forces are at work in determining what types of contracts people *consider* to be standard.

To get a handle on this issue, we model the bargaining process as a pure coordination game. Each party demands certain terms in the contract. They enter into a contractual relationship if and only if their demands are consistent. Each player's expectation about what the other will accept is shaped by precedent, that is, by the terms that other people have agreed to in similar situations. This feedback effect causes society to converge toward "standard contracts"—terms that are normal and customary in a given kind of relationship. When stochastic perturbations are introduced, the model makes predictions about the welfare properties of contracts that are most likely to be observed in the long run. In particular, it says that such contracts will tend to be *efficient*: no other terms offer higher expected payoffs for both sides. Furthermore, they will tend to be *fair* in the sense that the expected payoffs are more or less centrally located within the payoff opportunity set. While similar predictions can be obtained in a world where agents employ rational principles of contract selection, no such assumption is made here. Rather, the outcome emerges (without anyone intending it) from the interaction of many myopic agents, each of whom is concerned only with maximizing his own welfare.