Journal of Economic Perspectives—Volume 16, Number 2—Spring 2002—Pages 47–66

Evolution and Game Theory

Larry Samuelson

Introduced by John von Neumann and Oskar Morgenstern (1944), energized by the addition of John Nash's (1950) equilibrium concept, and popularized by the strategic revolution of the 1980s, noncooperative game theory has become a standard tool in economics. In the process, attention has increasingly been focused on game theory's conceptual foundations. Two questions have taken center stage: Should we expect Nash equilibrium play—that is, should we expect the choice of each player to be a best response to the choices of the other players? If so, which of the multiple Nash equilibria that arise in many games should we expect?

In the 1980s, game theorists addressed these questions with models based on the assumptions that players are perfectly rational and have common knowledge of this rationality. In the 1990s, however, emphasis has shifted away from rationality-based to evolutionary models. One reason for this shift was frustration with the limitations of rationality-based models. These models readily motivated one of the requirements of Nash equilibrium, that players choose best responses to their beliefs about others' behavior, but less readily provided the second requirement, that these beliefs be correct. Simultaneously, rationality-based criteria for choosing among Nash equilibria produced alternative "equilibrium refinements"—strengthenings of the Nash equilibrium concept designed to exclude implausible Nash equilibria—with sufficient abandon as to prompt despair at the thought of ever choosing one as the "right" concept. A second reason for the shift away from rationality-based game theory was a change in the underlying view of what games represent. It was once typical to interpret a game as a literal description of an idealized interaction, in which an assumption of perfect rationality appeared quite

■ *Larry Samuelson is Professor of Economics, University of Wisconsin, Madison, Wisconsin. His e-mail address is ⟨LarrySam@ssc.wisc.edu⟩.*

natural. It is now more common to interpret a game, like other economic models, as an approximation of an actual interaction, in which perfect rationality seems less appropriate.

The term "evolutionary game theory" covers a wide variety of models. The common theme is a dynamic process describing how players adapt their behavior over the course of repeated plays of a game. The interpretations of this dynamic stretch from biological processes acting over millions of years to cultural processes acting over generations to individual learning processes acting over the few minutes that pass between rounds of an experiment. The dynamic process potentially provides the coordination device that brings beliefs into line with behavior, providing the second requirement for a Nash equilibrium. It also provides a context for play that may be useful in assessing multiple equilibria.[1] Overall, it brings game theory closer to economics by viewing equilibrium as the outcome of an adjustment process rather than something that simply springs into being.

This essay first describes the basic techniques of evolutionary game theory. I then turn to the questions of whether evolutionary game theory provides support for equilibrium play and whether it provides insight into which equilibrium we might expect to see. Finally, it is useful to recall a time when general equilibrium theory was the new technique sweeping the profession, spurred by elegant proofs of existence and optimality and prompting concern about questions as to why we should expect equilibrium outcomes. The result was an explosion of work on tâtonnemont and other adjustment processes. This work taught us much about competitive equilibrium, but had little impact on the practice of economics. Will evolutionary game theory have a similarly negligible effect on what economists do? Perhaps not. Game theory has given rise to an equilibrium selection problem, potentially amenable to evolutionary techniques, that had no parallel in the case of general equilibrium theory. The key to the success of evolutionary game theory will be delivering results in this area that will change the way economists practice their craft. The final section suggests where we might look for such results.

Evolutionary game theory has been the subject of several surveys and texts, including Fudenberg and Levine (1998), Hofbauer and Sigmund (1988), Mailath (1998), Samuelson (1997), van Damme (1991, chapter 9), Vega-Redondo (1996), Weibüll (1995) and Young (1998). I will accordingly not attempt a survey of the literature in this paper and will also feel free to suppress technical details or simplify models whenever helpful.

[1] Dynamic models based on adaptive behavior have a long history in economics. A distinguishing feature of evolutionary game theory is an explicit model of the strategic considerations that give rise to individual behavior and that convert this behavior into economic outcomes.

Evolutionary Models

Biological Antecedents

True to its name, evolutionary game theory appeared first in biology. The central concept of an *evolutionarily stable strategy* was introduced by Maynard Smith and Price (1973) and developed further in Maynard Smith's (1982) influential *Evolution and the Theory of Games.* Dawkins (1989, p. 84) suggests that evolutionary stability is potentially "one of the most important advances in evolutionary theory since Darwin."

The context used to interpret an evolutionarily stable strategy envisions a large population of agents who are repeatedly, randomly matched in pairs to play a game. This underlying game is assumed to be symmetric, in the sense that i) the players choose their strategies from identical sets, and the payoff to a player choosing a particular strategy against an opponent choosing an alternative is the same regardless of the identities or characteristics of the players; and ii) players cannot make their strategy choices conditional on any characteristics such as which is the larger or older, or which is the row player. The payoffs in the game are assumed to represent "fitnesses," in the sense that a process of natural selection will favor those who earn higher payoffs.

Now suppose that everyone in the population plays a "common" strategy, except for a tiny toehold of "mutants" who play an alternative strategy. If the common strategy earns a higher expected payoff than the mutant strategy, then we can expect selection to eliminate the latter. If this outcome holds for any possible mutant strategy, then the common strategy is said to be evolutionarily stable. An evolutionarily stable strategy is thus a strategy that, once pervasive in the population, can repel any (sufficiently small) mutant advance.

Translating this higher-expected-payoff-than-any-mutant condition into payoffs, any strategy that is a strict best response to itself (that is, earns a strictly higher payoff against itself than does any alternative) will be evolutionarily stable. Because such a strategy earns a higher payoff against itself than does any mutant, it will earn a higher average expected payoff than the mutant in any population in which the mutant toehold is sufficiently small.

A strategy may also be evolutionarily stable if it is only a weak Nash equilibrium (that is, if there is some other strategy that fares as well against the candidate for evolutionary stability as does the candidate itself), but only if the candidate strategy then satisfies the stability condition that it earns a higher payoff when facing any alternative best response than does the alternative itself. In this case, the evolutionarily stable strategy secures a higher expected payoff in the population not through its unrivaled performance against itself, where it ties with the mutant, but by performing better against the mutant.

The criteria for evolutionary stability are thus more demanding than those for a Nash equilibrium. Any evolutionarily stable strategy must be at least a weak best response to itself, and is hence a Nash equilibrium, but a weak Nash equilibrium that fails the stability condition is not evolutionarily stable.

Figure 1
Hawk-Dove Game

	Hawk	Dove
Hawk	$\frac{1}{2}(V-C), \frac{1}{2}(V-C)$	$V, 0$
Dove	$0, V$	$\frac{V}{2}, \frac{V}{2}$

Maynard Smith (1982) opened his book with the hawk-dove game, which has since become a standard setting for discussions of evolutionary stability in biology. This game, shown in Figure 1, involves two players who contest a resource worth V. If one player is aggressive (Hawk) and the other acquiescent (Dove), then the former gets the resource and the latter nothing. Each has an equal chance at the resource if both are aggressive or both passive, with mutual aggression causing each to incur an injury cost of $C > V$ with probability $\frac{1}{2}$. The hawk-dove game has a unique evolutionarily stable strategy, given by the mixed strategy in which Hawk is played with probability V/C.[2]

Knowing that a strategy is evolutionarily stable tells us something about a population in which everyone chooses that strategy. But do we have any reason to expect such a state of affairs to arise? In response to this question, biologists have studied the population dynamics that lie behind the evolutionary stability concept more explicitly. In keeping with the interpretation of payoffs as fitnesses, let payoffs identify rates of reproduction. Then the composition of the population is described by the *replicator* dynamic in which the share of the agents playing a given strategy grows at a rate equal to the difference between the average payoff of that strategy and the average payoff of the population as a whole.[3] An evolutionarily stable strategy is asymptotically stable under the replicator dynamics, meaning that the dynamics converge to the evolutionarily stable strategy from all nearby population configurations, providing a dynamic motivation for the concept of evolutionary stability. In the hawk-dove game, for example, the replicator dynamic will converge to the state in which proportion V/C of the population plays Hawk. This reproduces the evolutionarily stable strategy, but in the form of a population in which V/C of the players in the population play Hawk and $1 - V/C$ play Dove, rather than a situation in which every player chooses a mixed strategy, which leads to Hawk with probability V/C.[4]

[2] The symmetry assumption rules out strategies such as "play Hawk when row player; Dove when column player." The mixed equilibrium must have the property that Hawk and Dove give identical expected payoffs against an opponent playing the mixed equilibrium. Hence, if p_H is the probability attached to Hawk, it must be that $p_H(\frac{1}{2} - (V - C)) + (1 - p_H)V = (1 - p_H)V/2$, which requires $p_H = V/C$.

[3] Let x_i be the share of the population choosing pure strategy i. The growth rate of the population share x_i is given by the difference between the average payoff of strategy i (denoted by π_i) and the average payoff of all strategies in the population (denoted by $\bar{\pi}$): $(dx_i/dt)(1/x_i) = \pi_i - \bar{\pi}$.

[4] Mixed-strategy equilibria have long been a source of uneasiness in game theory. Why should players who are indifferent between strategies, as they must be in a mixed equilibrium, randomly choose

The evolutionarily stable strategy of the hawk-dove game is also its unique symmetric Nash equilibrium. In many biological applications, the Nash equilibrium condition alone suffices to yield the desired outcome. As a result, perhaps the primary effect of the evolutionary stability concept in biology has been to popularize the ideas of a noncooperative game and Nash equilibrium.

The concepts of evolutionary stability and the replicator dynamics sweep away much that is of interest to biologists, most noticeably considerations arising out of the genetics of sexual reproduction. This neglect has prompted a continuing effort to embed evolutionary game theory in more realistic biological models (for example, Eshel, 1991; Eshel, Feldman and Bergman, 1998).

Evolutionary Stability in Economic Models

Evolutionary ideas have a long history in economics, with origins that predate biological applications. Darwin (1887, p. 83) acknowledged the influence of Malthus and the classical economists in the formation of his theory of natural selection. Alchian (1950) and Friedman (1953) popularized evolutionary metaphors to motivate the "as if" approach to optimization.[5] Economic theory is now routinely described as assuming not that people are relentless maximizers, but rather that some process of selection—perhaps the tendency of unprofitable firms to fail or the tendency of people to imitate their more successful counterparts—will cause us to observe people who act as if they are maximizing. This view allows optimization to be a tiny subset of the vast repertoire of possible human behavior, but makes it quite likely that the behavior we observe will be drawn from this subset.

Evolutionary game theory brings the evolutionary portion of these arguments out of the background. This approach initially gained popularity on the strength of evolutionary stability's ability to reject some seemingly implausible Nash equilibria. Consider the joint venture coordination game shown in Figure 2. Think of this as a case in which two players have the opportunity to form a joint venture that will earn a profit of 2 for each of them if they both choose In. If at least one player chooses Out, the opportunity dissipates with no reward and no cost to either player.

This game has two Nash equilibria, given by (In, In) and (Out, Out), but the former appears to be overwhelmingly more compelling than the latter. The intuition directing our attention to (In, In) is reproduced in the argument that only In is an evolutionarily stable strategy in this game. Because In is a best response to Out, and a superior response to itself, a population in which the joint venture is assiduously ignored could be invaded by mutants who exploit the venture, losing

between these strategies in precisely the proportion required for equilibrium? Harsanyi (1973) introduced a model in which players' payoffs are those specified in the game, plus small privately observed perturbations. Players choose pure strategies that are strict best responses given their payoff perturbation. Remarkably, Harsanyi showed that no matter what the nature of the perturbations, the resulting game has a *pure-strategy* equilibrium that approximates the original mixed equilibrium of the unperturbed game. The evolutionary model of a polymorphic population in which some players choose Hawk and some Dove similarly "purifies" mixed equilibria.

[5] The paper by Nelson and Winter in this symposium discusses this "as if" approach in greater detail.

Figure 2
Joint Venture Game

	In	Out
In	2, 2	0, 0
Out	0, 0	0, 0

nothing against the those who ignore the opportunity and gaining when encountering one another, thus ensuring that they fare better on average than those who play Out. Evolutionary stability thus directs our attention to the more plausible Nash equilibrium.

One could also justify equilibrium (In, In) by appealing to the cornerstone of the equilibrium refinements literature of the 1980s, that players should avoid weakly dominated strategies. Strategy In is the only undominated strategy in this game. Selten (1975) introduced the concept of a perfect equilibrium to capture the sense in which dominance considerations make (In, In) more appealing than (Out, Out) in Figure 2. The intuition behind perfect equilibria is that there is always some chance that *any* strategy might be played, perhaps by mistake or through some environmental tremble, and so one should protect oneself against the unexpected by avoiding dominated strategies. The concept of a proper equilibrium (Myerson, 1978) strengthens this by assuming that trembles discriminate among inferior strategies, attaching arbitrarily less probability to those that are more inferior. But why should rational players tremble, and why should mistakes or trembles have any relationship to payoffs? Evolutionary stability provides an answer: In symmetric games, evolutionarily stable strategies induce equilibria that are proper (and hence perfect) (van Damme, 1991, Theorem 9.3.4). Any strategy that is chosen without regard to trembles can thus be displaced by an evolutionarily superior mutant.

Unfortunately, complications loom close behind that prevent us from simply ending the argument with the assertion that evolutionary stability implies proper equilibrium. Some anomalies arise out of the fact that not all proper equilibria are evolutionarily stable. More importantly, the convenient link between evolutionary stability and refinements of the Nash equilibrium concept does not extend beyond symmetric games. Instead, Selten (1980) has shown that in asymmetric games, a strategy is evolutionarily stable only if it is a *strict* Nash equilibrium for all players to choose this strategy, that is, only if each player's strategy is a unique best response to the strategies of the other players.

This result has striking implications. In an extensive-form game, for example, a strategy can be evolutionarily stable only if it is a pure strategy and causes every contingency in the extensive form to be realized in equilibrium. Suppose instead that some contingency is missed. Then there must be an alternative best response whose behavior differs from that of the candidate strategy only in circumstances that are never realized. Because the differences in the strategies are never realized in the course of playing the game, both strategies attain identical payoffs, leaving no opportunity for the candidate strategy to gain a payoff advantage against the alternative and hence no prospect of expelling the latter from the population. The candidate strategy then is not evolutionarily stable.

Do we really care if a strategy fails to be evolutionarily stable because there are mutants whose differences do not appear in the course of play? Why not simply revise the definition of evolutionary stability, still requiring that no mutant earn a higher expected payoff than the evolutionarily stable strategy, but allowing the possibility of a mutant whose play duplicates that of the candidate strategy? Doing so yields what Maynard Smith (1982, p. 107) called a *neutrally stable strategy.*[6] Let the mutants come, as long as they produce identical behavior.

Unfortunately, mutants who produce behavior identical to that of an existing strategy can be of tremendous importance. For example, the *Tit-for-tat* strategy, which cooperates on its first move and mimics the opponent's previous move thereafter, was initially suggested as an evolutionarily stable strategy for the repeated prisoners' dilemma, a result made all the more appealing by the fact that it yields perpetual cooperation. However, the strategy Cooperate, that simply cooperates all of the time regardless of what its opponent does, behaves identically to Tit-for-tat under any circumstances that can be reached when the two play the game, differing only in the event of the unrealized contingency of a defection.[7] Unlike the Tit-for-tat strategy, the Cooperate strategy can be exploited by opponents who defect. The stability of cooperative behavior can thus depend critically upon whether a population consists primarily of Tit-for-tat or the seemingly identical Cooperate.

The potential instability of neutrally stable strategies is a general problem. For example, no strategy can gain a payoff advantage over the strategy of Tat-for-tit, one of the neutrally stable strategies in Binmore and Samuelson's (1992) repeated prisoners' dilemma with complexity costs.[8] But there are alternative best responses that do just as well as Tat-for-tit and hence that can creep into the population without being expelled by Tat-for-tit. In addition, there are predatory strategies that can gain an advantage over some of these infiltrators, so that the appearance of the

[6] The concept of neutral stability has proven useful. Binmore and Samuelson (1992), working with repeated games (including the prisoners' dilemma), in which players prefer simpler to more complex strategies (other things equal), show that neutrally stable strategies exist and give efficient outcomes. Fudenberg and Maskin (1990) obtain a similar result. Matsui (1991) and Kim and Sobel (1992) exploit analogous ideas to identify conditions under which evolution will select efficient outcomes in cheap-talk games, which are games in which play is preceded by opportunities for nonbinding communication, or "cheap talk." (For an introduction to cheap-talk games in this journal, see Farrell and Rabin, 1996.) These results are welcome in light of the frustrating abundance of equilibria in both repeated games and cheap-talk games.

[7] Hence, Tit-for-tat is neither a strict Nash equilibrium nor evolutionarily stable (nor is any other strategy) in the repeated prisoners' dilemma.

[8] The Tat-for-tit strategy defects in the first period and thereafter changes its action whenever the opponent defects. Two Tat-for-tit players produce an initial period of defection followed by perpetual cooperation, enforced by the fact that a defection prompts the opponent to switch back to defecting. Once complexity costs are incorporated, Tit-for-tat is not even neutrally stable in the repeated prisoners' dilemma, with the simpler Cooperate strategy being a better response.

latter opens the door to mutants who may destroy the efficient outcome. There seems to be no hope for ascertaining whether Tat-for-tit or efficient play more generally can be expected to survive such onslaughts short of explicitly examining the evolutionary dynamics. As a result, attention in evolutionary game theory has increasingly turned to dynamic models.

Evolutionary Dynamics in Economic Models

Economists working with dynamic models envision a population of players who are repeatedly, randomly matched to play a game. Each player is equipped with a behavioral rule that chooses strategies in the game as a result of the player's experience, typically interpreted as modeling a learning or imitation process. A wide range of such behavioral rules have been examined, from simple stimulus-response rules that attribute no cognitive activity to the agent to models in which agents play best responses to expectations that are the result of complicated Bayesian inference problems, though it is common to build some degree of "bounded rationality" into the model. If nothing else, players are typically assumed to ignore any effect that their current actions might have on the future behavior of their compatriots, a formulation often motivated by an assumption that the population is quite large.

This setting immediately suggests some limitations for evolutionary game theory. First, if bounded rationality is an important constraint, then we would expect subjects simply to ignore those games in which the payoff consequences of their choices are small. Scarce reasoning resources will not be expended without some reasonable prospect of gain. Secondly, if players must learn which strategies are advantageous, then we cannot expect to say much about games that are played too infrequently or that are too complicated for such learning to occur. Combining these considerations, we might expect the amount of experience required for players to hit upon suitable behavior to decrease when the stakes increase and the complexity of the problem decreases, as players juggle the bounded rationality constraint by bringing more sophisticated learning processes to bear on games where they are more likely to make a difference. Very few people play bridge well the first time they play it, no matter how much they have studied the game beforehand. No matter how often one allows them to try, and how much one pays for a correct answer, most people will not learn to prove Goldbach's Conjecture (that every even integer larger than two is the sum of two primes). Most of us continually fall prey to optical illusions, simply because it does not pay to be constantly on guard against them.

We thus cannot expect the results of evolutionary game theory always to be applicable. I do not find this constraint particularly troubling, nor do I think it special to games. Instead, I suspect that this is a general characteristic of human behavior: it becomes more deliberate, more measured and seemingly more rational as the consequences increase, the problem becomes more straightforward and familiarity with the problem increases. The question of how important and how familiar a problem must be before our analysis is likely to be useful pervades all of economics.

One approach to dynamic evolutionary models is based on examining deterministic difference or differential equations that describe the proportion of the players in the population playing the various strategies. The typical motivation behind such studies is that individual behavior is likely to be both quite complex and stochastic, but that forces akin to the law of large numbers are likely to ensure that in large populations this behavior averages to something reasonably deterministic and, one hopes, reasonably simple. The dynamics themselves come in many varieties. Some studies have simply adopted the replicator dynamics. However, in response to concerns that the biologically motivated replicator dynamics may not be appropriate in economic settings, work is typically done with a more general dynamic satisfying a monotonicity condition. This latter condition imposes some version of the requirement that the population shares of high-payoff strategies grow more quickly than those of low-payoff strategies, without imposing the specific structure of the replicator dynamics, and is often interpreted as assuming that, on average, the players are able to switch from worse to better strategies.[9]

Alternatively, evolutionary models have been built up from explicit specifications of individual behavior, their authors preferring to tackle the vagaries of individual choice directly rather than smoothing them out in an aggregate dynamic. Kandori, Mailath and Rob (1993) consider a collection of players who are repeatedly matched to play a coordination game such as that shown in Figure 3. (Foster and Young, 1990, and Young, 1993, work with similar models.) Time is measured in discrete periods. In each period, the agents are matched to play a round-robin tournament with the other agents in the population. With high probability, each player chooses the strategy that is a best response to the previous period's distribution of play. With small probability, the player is a "mutant" who chooses strategy X or Y, each with probability $\frac{1}{2}$. The result is a stochastic process, where the state space is the set of possible specifications of which players choose which strategies. Behavior in any single period will always be unpredictable, but over time, average behavior will converge to a stationary distribution. Kandori, Mailath and Rob (1993) and Young (1993) show that this type of model can yield particularly strong results when attention is directed to the "limiting distribution," which is obtained by examining the limit of the stationary distributions as the mutation rate becomes arbitrarily small.

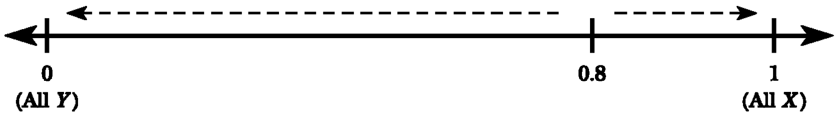
Figure 4 shows a phase diagram for a monotonic dynamic operating on the coordination game of Figure 3. The axis measures the proportion of the population playing strategy X, ranging from zero (corresponding to the equilibrium (Y,Y)) to one (corresponding to the equilibrium (X,X)). Whenever less than 80 percent of

[9] Considerable work has also been done with generalizations of fictitious play (for example, Fudenberg and Levine, 1998), in which players choose best responses to the average of their opponents' past play. This work reinterprets what was originally proposed as a technical tool for calculating Nash equilibria as a learning process.

Figure 3
Coordination Game

	X	*Y*
X	9, 9	0, 7
Y	7, 0	8, 8

Figure 4
Phase Diagram for a Monotonic Dynamic, Applied to the Coordination Game of Figure 3



the population initially plays *X*, so that *Y* is a best response, the game converges to the equilibrium in which everyone plays *Y*. If more than 80 percent initially play *X*, then *X* is a best response and the equilibrium in which everyone plays *X* is approached. The system thus always approaches an equilibrium, but which one is selected depends upon the happenstance of how the population is originally distributed between the two strategies.

In contrast, Kandori, Mailath and Rob (1993) conclude that, no matter what the initial condition, the system spends virtually all of its time at the equilibrium (Y,Y), for a sufficiently small mutation probability. Because almost all agents choose a best response in each period, the stationary distribution concentrates virtually all of its probability mass on states where almost all agents play *X* or almost all play *Y*. Once near such a state, the system tends to remain there. Occasionally, sufficiently many players will just happen to be mutants, and to switch strategies, as to switch the system from one in which *X* (or *Y*) is a best response to one in which *Y* (or *X*) is a best response, flipping the system to the other end of the state space. The phase diagram in Figure 4 shows that more mutations (80 percent of the population) are required to make *X* a best response in a population where everyone plays *Y* than to accomplish the reverse transition (requiring only 20 percent mutants), making the latter transition more likely. As the mutation probability gets small, the latter transition becomes arbitrarily more likely, causing all of the probability in the stationary distribution to accumulate on the equilibrium (Y,Y).

The two types of model thus appear to give disconcertingly different results. If most of the population initially plays strategy *X*, for example, then the deterministic dynamics predicts convergence to the equilibrium (X,X), while the stochastic process directs sole attention to the equilibrium (Y,Y).

Upon closer examination, it is not the models that differ so much as the questions we ask of the models. Binmore, Samuelson and Vaughan (1995), building on the work of Boylan (1995), show that both models provide approximations of the underlying stochastic process. A deterministic dynamic, whose details depend upon the nature of the individual learning processes, (approximately) describes the behavior of the system over finite periods of time, being applicable to longer and longer periods of time in the case of larger and larger populations.[10] At the same time, the limiting distribution describes the behavior in the limiting case of an infinite time horizon.

Results of this kind allow us to see that seemingly quite different evolutionary models are often compatible. The key question is not so much one of which model to select from a list of conflicting contenders, but rather which information to seek about a single underlying evolutionary process, with the answer implying the relevant model. Evolutionary game theory is thus unlikely to provide context-free answers, nor will it identify a single equilibrium concept as unquestionably "right." The relevant time horizon, population size, individual behavior specification and the interaction rules will all depend upon the setting to which the evolutionary analysis is applied and, hence, so will the specification of an appropriate model.

I view this abundance of possible outcomes as an advantage. Much of the difficulty in interpreting the contending equilibrium refinements of the 1980s appeared because the models were divorced from their context of application in an attempt to rely on nothing other than rationality. The resulting models contain insufficient information about the underlying strategic interaction to answer the relevant questions, at least if game theory is intended to model real interactions rather than to ponder philosophical points. Certain behavior may be quite likely in some settings and quite absurd in others. It is interesting that students seem to grasp this point instinctively—a common response when queried about how they would behave in a game is to ask, "Who is my opponent?" In general, the process by which people find their way to an equilibrium may be littered with accidents of history, framing effects, bandwagon effects and endogenous conventions. A useful theory must incorporate these considerations. Evolutionary game theory provides some tools for bringing them into the theory.

Why Equilibrium?

With this background in mind, we turn to our first question: Do evolutionary models give us any reason to choose Nash equilibrium as a solution concept? Given our previous discussion of evolutionary stability, attention naturally turns to dynamic models.

Consider first a stationary state of the replicator dynamic, describing the

[10] Surprisingly, in some cases, the replicator dynamic emerges as the relevant deterministic approximation from a model based on learning considerations that appears to have no biological connection, substituting imitation for reproduction as the driving force.

proportions of the population playing the game's various strategies. We say that this state is *stable* if small perturbations away from the stationary state proportions cannot give rise to dynamics that take the system far from these proportions. Instead, an initial condition close to the stationary state proportions ensures that the system remains perpetually close to these proportions.

If a state is to be stationary, then all of the strategies played by various members of the population must give the same payoffs, since otherwise the population proportion attached to high-payoff strategies would be growing at the expense of low-payoff strategies, vitiating stationarity. However, a stationary state may still not be a Nash equilibrium, because there may be superior replies that are not played by any member of the population, and hence whose population proportion cannot grow (via reproduction or imitation) from an initial proportion of zero. Once a perturbation moves the system to a nearby state in which such a superior strategy is played, the latter's population share will grow, leading the system away from the original stationary state and ensuring that the latter is not stable. A stationary state can thus be stable only if there are no superior best responses, in which case the stationary state must correspond to a Nash equilibrium.

Results of this form—that stability implies Nash equilibrium—also emerge from the various monotonic dynamics that generalize the replicator dynamic, as well as from a wide variety of stochastic models based on individual behavior (with suitably modified notions of stability). Such results are sufficiently common that a typical reaction to an evolutionary model featuring convergence to an outcome that is *not* a Nash equilibrium would be to argue that the model is misspecified or implausible.

We thus have a qualified positive answer to the question of whether evolutionary game theory provides a motivation for Nash equilibrium. I characterize this as a qualified positive answer because of the conditional nature of the result: an outcome is a Nash equilibrium *if it is stable.*

Some game theorists would have preferred a result of the form that an evolutionary process *must* produce convergence to a Nash equilibrium. However, there are ample examples of simple evolutionary models that yield cyclic or even chaotic behavior instead of convergence to a Nash equilibrium. It remains an open question whether any plausible evolutionary process exists that invariably ensures convergence.[11] But even if such a process existed, we have little reason to believe that it would faithfully reflect actual behavior.

I view the "stability implies Nash" result as putting game theory on much the same footing as the rest of economics. We do not believe that markets are always in equilibrium, just as we do not believe that people are always rational or that firms always maximize profits. But the bulk of our attention is devoted to equilibrium

[11] Hart and Mas-Collel (2000) present a process that always converges to a correlated equilibrium, but argue that the extra complexity of the set of Nash equilibria and the natural propensity for the history of an adaptive process to induce correlation suggest that we cannot expect the process always to arrive at a Nash equilibrium.

models either because we hope that equilibrium behavior is sufficiently persistent and disequilibrium behavior sufficiently transient that behavior that is robust enough to be an object of study is (approximately) equilibrium behavior, or because studying equilibrium behavior is our best hope for gaining insight into more ephemeral disequilibrium behavior. Evolutionary game theory thus provides little reason to believe that equilibrium behavior should characterize all games in all circumstances. But it provides reason to hope that behavior that comes into our field of study is likely to be equilibrium behavior. In this sense, we obtain a stronger motivation for Nash equilibrium than that provided by rationality-based models.

Which Equilibrium?

Does evolutionary game theory direct our attention to some Nash equilibria rather than others? Again, previous discussion directs our attention to dynamic models. The results are surprising in the extent to which they do both more and less than traditional equilibrium refinements.

Choosing Between Strict Nash Equilibria

To see how evolutionary game theory does more than traditional refinements, consider again the coordination game of Figure 3. This game has two pure-strategy equilibria, given by (X, X) and (Y, Y), each of which is strict, in the sense that each player has a unique best response. With the notable exception of Harsanyi and Selten (1988), the equilibrium refinements literature has concentrated on eliminating Nash equilibria in which there are alternative best responses. Strict Nash equilibria survive all of the conventional refinements, reflecting an intuition that a situation in which everyone has a strict incentive to maintain their current behavior is not easily destabilized.

In contrast, the evolutionary models of Young (1993) and Kandori, Mailath and Rob (1993) make a distinction between these two equilibria, with the limiting distribution allocating almost all of its probability to equilibrium (Y, Y) in Figure 3. More generally, these models select the equilibrium in 2 × 2 games with the larger basin of attraction under a monotonic dynamic.[12]

The finding of Kandori, Mailath and Rob (1993) and Young (1993) that the equilibrium with the larger basin of attraction is selected can be reversed with appropriate modifications to the model, as in Robson and Vega-Redondo (1996). The important result is not the selection of a particular equilibrium, but rather the

[12] The basin of attraction under a monotonic dynamic is the set of (possibly mixed) strategy profiles to which the equilibrium strategy in question is a best response. In 2 × 2 games, an easy test of which equilibrium has the largest basin of attraction is provided by the fact that this equilibrium is the best response to a 50:50 mixture between the two strategies. This is the equilibrium that Harsanyi and Selten (1988, pp. 82–84) refer to as risk dominant.

departure from the bulk of the equilibrium refinements literature in distinguishing between contending strict Nash equilibria.

This ability is not an unqualified success. The limiting stationary distribution, as the probability of a mutation goes to zero, may be a reasonable approximation only of what occurs after long periods of time, with this waiting time becoming very long when the probability of a mutation is quite small. In some cases, the implied waiting times will be sufficiently long that our interest will center on shorter horizons and the stationary distribution will be irrelevant. In other cases, the stationary distribution may be more useful. Beginning with the work of Ellison (1993), it has been recognized that waiting times may be significantly reduced if there are spatial or "local" patterns to the interaction between agents. Young (1998) discusses the evolution of social structures that may occur over sufficiently long periods of time for the theory to be applicable. Much remains to be done, but it is clear that evolutionary game theory has provided new tools to address a difficult question.

Equilibrium Refinements

To see how evolutionary game theory does less than the equilibrium refinements literature, we return to the cornerstone of the refinements literature: the presumption that weakly dominated strategies should not be played. Consider the game whose normal and extensive forms are shown in Figure 5. Binmore, Gale and Samuelson (1995) interpret this as a simplified version of the ultimatum game. Player 1 must propose an amount of a surplus of size 4 to offer to player 2 and can choose either a high offer of 2 or a low offer of 1. A high offer is assumed to be accepted, while a low offer may be either accepted ("Yes") or rejected ("No").

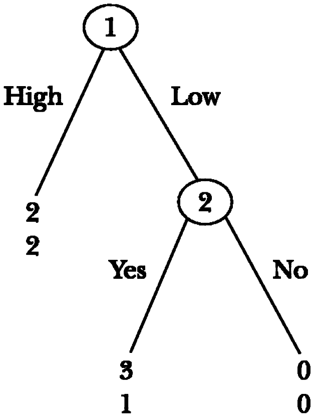
An equilibrium of this game is subgame perfect if player 1's choice is a best response to player 2's choice and if player 2's Yes/No decision would be a best response in the event that player 1 chooses Low. Backward induction identifies the only subgame perfect equilibrium of this game: player 2 accepts low offers, and as a result, player 1 makes a low offer. There are other Nash equilibria in which player 1 makes a high offer and player 2 plays No with probability at least 1/3.

No is a dominated strategy for player 2. It can never earn a higher payoff than Yes, and one would accordingly expect an evolutionary process to exert constant pressure against No. Suppose, however, that a large fraction of the player-2 population initially plays No. High will then produce a higher average payoff for player 1 than Low, and an evolutionary process will also exert pressure against Low. But as fewer and fewer player 1s choose Low, the payoff disadvantage of No dissipates, and hence, so does its evolutionary disadvantage. The result may be convergence to an outcome in which player 1 offers High and a significant fraction of player 2s would reject Low if offered. The dominated strategy No is thus not eliminated.

Binmore, Gale and Samuelson (1995) and Roth and Erev (1995) fill in the details of this argument. However, it seems as if this argument relies too heavily on the fact that player 2s' choice of No is never tested if player 1s make high offers. We expect the world to be a noisy place, certainly noisier than our simple models. It

Figure 5
Simplified Ultimatum Game

	Yes	No
High	2, 2	2, 2
Low	3, 1	0, 0



seems that this noise should ensure that Low never dies out of the population completely and, hence, that No is always inferior to Yes and should accordingly be eliminated from the population. Binmore and Samuelson (1999) show that this need not be the case and that the outcome depends delicately on the specification of noise. By continually injecting strategy Low into population 1, noise introduces a pressure against No, though this pressure may be quite weak if most of the population plays High. In addition, this same noise has the potential to introduce a counterpressure by continually injecting No into population 2. Either force may win, raising the possibility that dominated strategies are not eliminated from the population.

The conventional interpretation behind these dynamic models is that they represent a learning process. One response to the previous paragraph is then to argue that the ultimatum game is transparently simple, making it hard to imagine any learning at all being required. This is especially the case for the responders. Their task is to accept some money or decline it. Then what do they have to learn? Whether money is good? Much depends, however, on our remembering that the games with which we work are meant to be not literal descriptions of reality, but rather models of interactions that may appear to be much more complicated to the participants. We should not confuse our ability to find Yes transparently obvious in the model with a similar ability on the part of the player "on the ground." President John F. Kennedy is often characterized as having presented Premier Nikita Khrushchev with an ultimatum during the Cuban missile crisis of 1963. Is it likely that Khrushchev found the resulting choice as simple as that shown in Figure 5?

As the dominance relations become more complicated, the elimination of dominated strategies becomes all the more problematic. Whereas suitably tailored evolutionary models will eliminate dominated strategies in certain contexts (for example, Hart, 2002), a reasonable characterization of the combined results of the literature is that one generally cannot rely on evolutionary processes to eliminate weakly dominated strategies, much less to perform iterated eliminations. Given the close link between dominance and backward induction arguments, it is no surprise that evolutionary models also provide very little motivation for backward induction. In terms of sorting between Nash equilibria, the lesson of evolutionary game theory is that we should be less anxious to apply dominance or backward induction based refinements than the refinements literature would suggest.

This finding dovetails nicely with the recent experimental literature, which provides ample reason to believe that backward induction should not be taken for granted (Davis and Holt, 1993, chapter 5; Roth, 1995). But in the experiments, the game literally *is* as transparent as that shown in Figure 5. Then why do we suppose that responders require learning or trial-and-error or experience to know what to do? We must remember that while the game itself is transparent, the context in which it is played, including the absence of any chance of repeated play or breach of anonymity between the players, is quite foreign. What do players do when faced with an unprecedented context? One possibility is that they strip away the context and analyze the game. Another is that they search their experience for the closest analogies they can find, using the game's context as a clue in searching for a similar situation and choosing behavior they have found to be effective in the latter. It may then take considerable experience to hit upon appropriate analogies, producing a dynamic process that, for reasons described above, need not produce the backward induction solution. The role of analogies in reasoning is pursued further in Jehiel (2000) and Samuelson (2001).

What Do We Take Away?

Will evolutionary game theory have an impact on the way people practice game theory, or will it fade away, leaving economists to carry on as they have before? The latter will surely be its fate if it does nothing other than ease our consciences a bit when doing what we've been doing all along, namely examining Nash equilibria. But I believe that evolutionary game theory has the potential to do more.

Evolutionary game theory will have done much if we simply take seriously the caution that dominance and backward induction arguments are not as compelling as they may first appear. It is common in models of bargaining, contracting and exchange to assume that an agent can be pushed to the brink of indifference and still be relied upon to agree to the deal. Though these arguments appear in a variety of guises, they are all variations on the assertion that the subgame perfect equilibrium will appear in the ultimatum game. The more we learn from evolutionary games, the less certain can we be that we have good reason for doing so.

For evolutionary game theory to realize its potential, however, it must go beyond warnings about what we should not do to provide results concerning what we should do. Here, evolutionary game theory runs the risk that plagued the equilibrium refinements literature: so many equilibrium concepts and so little basis

Figure 6
Coordination Games

	X	*Y*
X	45, 45	0, 35
Y	35, 0	40, 40

	X	*Y*
X	45, 45	0, 40
Y	40, 0	20, 20

	X	*Y*
X	45, 45	0, 42
Y	42, 0	12, 12

for choosing among them in the abstract. However, I regard three areas of research as particularly promising.

The first is research that removes evolutionary game theory from its abstract setting and links the theory to observed behavior, either in the laboratory or in the field. For example, Battalio, Samuelson and van Huyck (2001) examine the three games shown in Figure 6. In many respects, these games are strategically identical. They have identical equilibria and best-response correspondences, and they have identical phase diagrams under best-response, replicator or monotonic dynamics with that phase diagram given by Figure 4. Many rationality-based models would accordingly treat these three games as equivalent. They differ, however, in that no matter what strategy one expects one's opponent to play, the premium on playing a best response (and hence the penalty for a suboptimal response) increases as one moves from right to left through the three games. If evolutionary models are on the mark in thinking of behavior as being shaped by trial-and-error learning, and in thinking of these processes as being more effective when it is more important to make good choices, then we would expect behavior to adjust to an equilibrium more rapidly as one moves from right to left. This is indeed the pattern in the data (Battalio, Samuelson and van Huyck, 2001, Figure 6 and Table 4).

This is only a single, small step, all the smaller because it examines a behavioral prediction that is particularly intuitive and that might also emerge from many other models. But more research is appearing that melds evolutionary models with behavioral observation. In assessing this work, one must realize that explaining the dynamics of individual behavior is a formidable task. The early steps will be modest, but will hold great promise.

Second, evolutionary models rely on settings in which games are played repeatedly. Literally speaking, we never face precisely the same decision twice. Instead, our hope must be that people face successive games that are sufficiently similar as to be viewed as essentially the same game. Again, this is consistent with a view of games as models, perhaps constructed by the players themselves, of more complicated strategic interactions.

This view of how people approach games not only potentially widens the purview of evolutionary game theory, but also provides some important insights into how people play games. Return to the question of which equilibria we should expect in the repeated prisoners' dilemma. The intuition that the players are likely to achieve mutual cooperation conflicts with the intuition that players should have a preference for simple strategies. In particular, the most common means by which cooperation is thought to be sustained is through strategies like Tit-for-tat, which begin play by cooperating and continue to cooperate as long as the opponent does so. In equilibrium, the ability to punish defections is never used. But this allows the players to simplify their strategies, at no sacrifice in payoffs, by eliminating the unused punishment capabilities. This leaves strategies that always cooperate, but whose vulnerability to defecting opponents precludes the existence of an equilibrium, threatening the ability to sustain mutual cooperation.

The conventional response is to seek strategies that cooperate nearly all of the time while still using their punishment capability (Abreu and Rubinstein, 1988; Binmore and Samuelson, 1992). Suppose, however, that we think of people as facing a variety of prisoners' dilemma situations. In some of these, the shadow of the future will be insufficiently important to induce cooperation, and mutual defection will be the fare. In others, the future will be important, and cooperation can be supported by a grim strategy that cooperates initially, doing so as long as the opponent does, switching to defection otherwise. Notice now that this strategy cannot be costlessly simplified by deleting the ability to defect, since this ability is needed to handle those situations in which defection is optimal. Considering the games together thus ensures that it is always important to be able to punish defection, allowing cooperation to be sustained whenever it is feasible without running afoul of complexity constraints.

This view of cooperation is reminiscent of work in evolutionary psychology, suggesting that people have a deep-seated ability to detect and to respond to cheating on norms of behavior (Cosmides and Tooby, 1992). The common theme is that this punishment ability is a general purpose capability applied as part of one's behavioral mix in a wide variety of settings. In some games, it may never be used, but the possibility that it might be eliminated in a quest for simplicity is squelched by its usefulness in other problems.

Finally, I think we have much to learn from pushing the evolutionary point of view beyond the simple question of how people behave in games. In particular, our evolutionary background has much to tell us about some of the idiosyncrasies of our preferences. Consider, for example, the question of why we have emotions. I suspect that emotions help us cope with our complex environment. An appropriately chosen rule of thumb of the form "do the fair thing" or "retaliate when crossed" may be optimal not because it allows us to commit in a seemingly irrational manner to things we would otherwise not do, but because it allows us to simplify the process by which we arrive at what we would otherwise want to do. This is especially the case if we think of people being faced with a vast variety of games that differ in intricately complex ways, which must then be simplified into something that can be tractably handled. Lumping possibly dissimilar games together, exploiting analogies between games and creating such analogies by labeling certain outcomes as fair may all be useful weapons in this process. In essence, we (or Nature, through the process of evolution) simplify our lives by deciding that things are fair because they are the things we do, not by doing things because they are fair. There may be much to be learned from such an extension of evolutionary game theory.

■ *I thank Ken Binmore, George Mailath, Brad De Long, Timothy Taylor and Michael Waldman for helpful comments and discussions and thank the National Science Foundation for financial support.*

References

Abreu, Dilip and Ariel Rubinstein. 1988. "The Structure of Nash Equilibrium in Repeated Games with Finite Automata." *Econometrica.* November, 56:6, pp. 1259–281.

Alchian, Armen. 1950. "Uncertainty, Evolution, and Economic Theory." *Journal of Political Economy.* 58, pp. 211–21.

Battalio, Raymond, Larry Samuelson and John van Huyck. 2001. "Optimization Incentives and Coordination Failure in Laboratory Stag Hunt Games." *Econometrica.* May, 69:3, pp. 749–64.

Binmore, Ken and Larry Samuelson. 1992. "Evolutionary Stability in Repeated Games Played by Finite Automata." *Journal of Economic Theory.* August, 57:2, pp. 278–305.

Binmore, Ken and Larry Samuelson. 1999. "Evolutionary Drift and Equilibrium Selection." *Review of Economic Studies.* April, 66:2, pp. 363–94.

Binmore, Ken, John Gale and Larry Samuelson. 1995. "Learning to be Imperfect: The Ultimatum Game." *Games and Economic Behavior.* January, 8:1, pp. 56–90.

Binmore, Ken, Larry Samuelson and Richard Vaughan. 1995. "Musical Chairs: Modeling Noisy Evolution." *Games and Economic Behavior.* October, 11:1, pp. 1–35.

Boylan, Richard T. 1995. "Continuous Approximation of Dynamical Systems with Randomly Matched Individuals." *Journal of Economic Theory.* August, 66:2, pp. 615–25.

Cosmides, Leda and John Tooby. 1992. "Cognitive Adaptations for Social Exchange," in *The Adapted Mind.* Jerome H. Barkow, Leda Cosmides and John Tooby, eds. Oxford: Oxford University Press, pp. 163–228.

Darwin, Charles. 1887. *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter, Second Edition, Volume 1.* Francis Darwin, ed. London: John Murray.

Davis, Douglas D. and Charles A. Holt. 1993. *Experimental Economics.* Princeton: Princeton University Press.

Dawkins, R. 1989. *The Selfish Gene.* Oxford: Oxford University Press.

Ellison, Glenn. 1993. "Learning, Local Interaction, and Coordination." *Econometrica.* September, 61:5, pp. 1047–072.

Eshel, Ilan. 1991. "Game Theory and Population Dynamics in Complex Genetical Systems: The Role of Sex in Short Term and in Long Term Evolution," in *Game Equilibrium Models.* Reinhard Selten, ed. Berlin: Springer-Verlag, pp. 6–28.

Eshel, Ilan, Marcus W. Feldman and Aviv Bergman. 1998. "Long-Term Evolution, Short-Term Evolution, and Population Genetic Theory." *Journal of Theoretical Biology.* 191:4, pp. 391–96.

Farrell, Joseph and Matthew Rabin. 1996. "Cheap Talk." *Journal of Economic Perspectives.* 10:3, pp. 103–18.

Foster, Dean and Peyton Young. 1990. "Stochastic Evolutionary Game Dynamics." *Journal of Theoretical Biology.* October, 38:8, pp. 219–32.

Friedman, Milton. 1953. *Essays in Positive Economics.* Chicago: University of Chicago Press.

Fudenberg, Drew and David K. Levine. 1998. *Theory of Learning in Games.* Cambridge: MIT Press.

Fudenberg, Drew and Eric Maskin. 1990. "Evolution and Cooperation in Noisy Repeated Games." *American Economic Review.* May, 80, pp. 274–79.

Harsanyi, John C. 1973. "Games with Randomly Distributed Payoffs: A New Rationale for

Mixed-Strategy Equilibrium Points." *International Journal of Game Theory.* 2, pp. 1–23.

Harsanyi, John C. and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge: MIT Press.

Hart, Sergiu. 2002. "Evolutionary Dynamics and Backward Induction." *Games and Economic Behavior.* Forthcoming.

Hart, Sergiu and Andreu Mas-Collel. 2000. "A Simple Adaptive Procedure Leading to Correlated Equilibrium." *Econometrica.* September, 68:5, pp. 1127–150.

Hofbauer, J. and K. Sigmund. 1988. *Evolutionary Games and Population Dynamics.* Cambridge: Cambridge University Press.

Jehiel, Phillippe. 2000. "Analogy-Based Expectation Equilibrium." Mimeo, University College London.

Kandori, Michihiro, George J. Mailath and Rafael Rob. 1993. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica.* January, 61:1, pp. 29–56.

Kim, Yong-Gwan and Joel Sobel. 1992. "An Evolutionary Approach to Pre-Play Communication." *Econometrica.* September, 63:5, pp. 1181–194.

Mailath. George J. 1998. "Do People Play Nash Equilibrium? Lessons from Evolutionary Game Theory." *Journal of Economic Literature.* September, 36:3, pp. 1347–374.

Matsui, Akihiko. 1991. "Cheap-Talk and Cooperation in Society." *Journal of Economic Theory.* August, 54:2, pp. 245–58.

Maynard Smith, John. 1982. *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.

Maynard Smith, John and G. R. Price. 1973. "The Logic of Animal Conflict." *Nature.* 246, pp. 15–18.

Myerson, Roger B. 1978. "Proper Equilibria." *InternationalJournal of Game Theory.* 7, pp. 73–80.

Nash, John F. 1950. "Equilibrium Points in n-Person Games." *Proceedings of the National Academy of Sciences.* 36, pp. 48–49.

Robson, Arthur J. and Fernando Vega-Redondo. 1996. "Efficient Equilibrium Selection in Evolutionary Games with Random Matching." *Journal of Economic Theory.* July, 70:1, pp. 65–92.

Roth, Alvin E. 1995. "Bargaining Experiments," In *Handbook of Experimental Economics.* John Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press, pp. 253–348.

Roth, Alvin E. and Ido Erev. 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior.* January, 8:1, pp. 164–212.

Samuelson, Larry. 1997. *Evolutionary Games and Equilibrium Selection.* Cambridge: MIT Press.

Samuelson, Larry. 2001. "Analogies, Adaptation, and Anomalies." *Journal of Economic Theory.* April, 97:2, pp. 320–66.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive-Form Games." *International Journal of Game Theory.* 4, pp. 25–55.

Selten, Reinhard. 1980. "A Note on Evolutionarily Stable Strategies in Asymmetric Animal Contests." *Journal of Theoretical Biology.* 84, pp. 93–101.

van Damme, Eric. 1991. *Stability and Perfection of Nash Equilibria.* Berlin: Springer-Verlag.

Vega-Redondo, Fernando. 1996. *Evolution, Games, and Economic Behavior.* Oxford: Oxford University Press.

von Neumann, John and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princeton : Princeton University Press.

Weibüll, Jurgen. 1995. *Evolutionary Game Theory.* Cambridge: MIT Press.

Young, Peyton. 1993. "The Evolution of Conventions." *Econometrica.* January, 61:1, pp. 57–84.

Young, Peyton. 1998. *Individual Strategy and Social Structure.* Princeton: Princeton University Press.