The Evolution of Norms[1]

Jonathan Bendor
Stanford University

Piotr Swistak
University of Maryland, College Park

Social norms that induce us to reward or punish people not for what they did to us but for what they did to other members of one's group have long been thought as sine qua non sociological and thus impossible to explain in terms of rational choice. This article shows how social norms can be deductively derived from principles of (boundedly) rational choice as mechanisms that are necessary to stabilize behaviors in a large class of evolutionary games.

INTRODUCTION

The question, Why are there norms? is perhaps one of the most fundamental problems that the social sciences have ever tackled. Its significance turns on the causal significance of its subject matter: "No concept is invoked more often by social scientists in the explanation of human behavior than 'norm'" (Gibbs 1968, p. 212).[2]

[1] For comments on earlier drafts of this article we would like to thank Jenna Bednar, Phillip Bonacich, Charles Cameron, Joe Harrington, Robert K. Merton, Harvey Molotch, John Padgett, Scott Page, Janek Poleszczuk, David Snow, and seminar participants at the University of California, Berkeley, Boston University, University of California, Irvine, Jachranka, Stanford University, University of California, Los Angeles, and the University of Chicago. Swistak thanks the Russell Sage Foundation, where he was a visiting scholar while much of the work on this article was done. Bendor thanks the Center for Advanced Study in the Behavioral Sciences for its financial and intellectual support. Direct correspondence to Jonathan Bendor, Graduate School of Business, Stanford University, 518 Memorial Way, Stanford, California 94305-5015. E-mail: bendor_jonathan@gsb.stanford.edu

[2] This quote was brought to our attention by Christine Horne (2001). A recent review of the literature in sociology, economics, and game theory on the emergence of norms is given in a series of papers by Horne (2001), Eggertsson (2001), and Voss (2001).

0002-9602/2001/10606-0001$02.50

The Problem

In the most general sense a norm might be regarded as any rule of behavior. Yet sweeping notions are rarely analytically useful. For this reason the concept of a norm has been typically restricted to a subset of behavioral rules. A norm, for instance, has often been considered to be in place "if any departure of real behavior from the norm is followed by some punishment" (Homans 1950, p. 123; see also Blake and Davis 1964, p. 457). Indeed if violating a rule never triggers any kind of punishment, calling it a norm would violate a common sociological intuition.[3] Thus, following Homans and Blake-Davis, we will understand norms as behavioral rules that are backed by sanctions. Hence when we ask, Why are there norms? it is norms in this sense that we have in mind. And as we shall later see, one of our results confirms the intuition that behavioral rules without sanctions are unimportant.

A second aspect of norms that seems fundamental to the notion relates to the nature of sanctioning. If a norm is violated, the obligation to impose punishment need not be restricted to those who were hurt by the original transgression; it can be extended to *third parties*—people unaffected by the deviation but who are in a position to punish the deviant. We will refer to rules of behavior that include third-party sanctions as *social norms*. Note that norms that obligate third parties to impose sanctions can be considered quintessentially *social*: by imposing requirements on an entire community and not merely on the interested parties, they create a general code of conduct. Violations of a general code matter to everyone in a community; they are not merely private matters between two parties. And because they are general, potentially embroiling people in conflicts that were not of their own making, it is precisely social norms that have been thought to be sine qua non sociological and thus impossible to explain in terms of rational choice. (Why, after all, should A punish B for what B did to C?) As we shall prove below, social norms can not only be derived as rational forms of behavior but, more important, they turn out to be *necessary* to stabilize behavior in groups and institutions. Consequently, most of our results focus on explaining social norms: their emergence and their features.

[3] If violating a rule *never* triggers *any* kind of punishment, a rule is completely discretionary and has no binding force. Indeed, even a common understanding of the term "norm" implies the existence of an enforcement mechanism. For example, the first definition of "norm" in Webster's *Ninth Collegiate Dictionary* is "a principle of right action binding upon the members of a group and serving to guide, control or regulate proper and acceptable behavior" (1989, p. 806).

Problems with the Existing Explanations

Given the fundamental importance of the problem of norms, it is not surprising that all of the social sciences have tried to resolve it. While the answers have crossed interdisciplinary boundaries, they have mainly fallen into two classes of explanations: one comes from the tradition of methodological individualism and economics, the other one draws on the structural-functional tradition of sociology. The first type of explanation assumes an individual to be a *Homo economicus* and explains actions in terms of individual choices and objectives. In the structural-functional tradition an individual is seen as *Homo sociologicus* and his or her behavior is explained in terms of group influence (e.g., conformity pressures) and the function that this behavior serves a group or society. Both modes of explanation have been criticized for some inherent problems; both have left a number of important questions unanswered. Before we present our solution to the problem of norms and the underlying theory, it will be good to remind the reader why the existing explanations have been criticized. The ensuing list of problems will set a reference point for the construction that follows. It will also serve as a set of objectives or criteria that can be used to evaluate the explanation proposed in this article.

Arguably, the most prominent answer to the problem of norms is that of functionalism and the homo sociologicus tradition: norms exist because they are functional for the group.[4] As is well known, however, a variety of problems attend this answer. First, there are conceptual issues: in particular, how are we to define the key concept, "functional"? Second, there are empirical problems: the strong functionalist thesis—all norms are functional for their collectivities—seems to founder empirically in the face of codes involving, for example, revenge and vendettas. Third, there are theoretical issues. As Stinchcombe (1968) and Elster (1983, 1989*b*) have argued, a complete functionalist argument must specify a selective pressure for the actions, rules, or structures that fulfill the functions. The selective pressure may be feedback loops (e.g., firms that do not maximize profits fail more often than those that do) or incentives (e.g., pressures to conform to roles or the internalization of values and norms) that directly

[4] This is what Wilbert Moore has called "the 'canonical' view that any observed cultural form or pattern of behavior must fit the system—that is, must have a function" (1978, p. 328). By functionalism we mean the school of thought that flourished in sociology in the 1950s and 1960s under the influence of Talcott Parsons and Robert Merton. The essential aspects of this paradigm go back to the early functionalism of cultural anthropologists such as Bronislaw Malinowski and Alfred Radcliffe-Brown as well as the tradition of European sociology, in particular that of Émile Durkheim.

induce decision makers to take the required actions.[5] Without some such specification the functionalist thesis becomes an illegitimate teleology. Fourth, structural-functional explanations of norms are generally static, a feature they share with most microeconomic and game-theoretic models of norms. The common problem with these theories is that while they can explain properties of equilibria, they cannot explain how a particular equilibrium has been arrived at. Thus these are static theories of inherently dynamic phenomena.

The fifth and perhaps the most important problem concerning functionalism (and a good part of standard sociological theory as well) was characterized by Coleman as follows: "Much sociological theory takes norms as given and proceeds to examine individual behavior or the behavior of social systems when norms exist. Yet to do this without raising at some point the question of why and how norms come into existence is to forsake the more important sociological problem in order to address the less important" (1990*a*, p. 241).

Coleman has seen the problem of deriving macro properties (properties of groups) from micro properties (properties of individuals) as the main unsolved problem of sociology and the program (Coleman 1986) for sociology's future development. "The emergence of norms is in some respect a prototypical micro-to-macro transition, because the process must arise from individual actions yet a norm itself is a system-level property" (Coleman 1990*a*, p. 244). For Coleman an effective theory of norms would begin with a set of assumptions about individuals in a group and conclude with a deductive derivation of group norms.

The Proposed Solution

In this article we show how evolutionary game theory can be used to explain an essential type of norms—social ones. In addition to bridging the micro-macro gap, this solution avoids, we believe, all the other problems of standard functional explanations that were mentioned above. Further, because we employ *evolutionary* game theory rather than (Nash) equilibrium analysis, we avoid the problems of static analyses that have attended efforts to explain norms via classical (noncooperative) game theory.

More specifically our answer to the question, Why are there norms? is an evolutionary one (Schotter 1981; Opp 1982; Sugden 1989; Bicchieri

[5] Parsons and other structural-functionalists recognized that individuals had to be motivated, somehow, to implement societal functions. But their reliance on the micro-mechanisms of conformity and internalization has been criticized for reflecting, as Wrong put it (1961), an "oversocialized conception of man."

1993, chap. 6; Binmore and Samuelson 1994). As Sugden put it, norms "have evolved because they are more successful at replicating themselves than other patterns [of behavior]" (1989, p. 97).[6] These replication processes are realized via incentives that induce members of a group to favor norms over other kinds of rules that do not encode norms. Since evolutionary game theory assumes individuals who, though boundedly rational, pursue norms that give them higher utilities, the main problem of a functional theory—How are functions realized?—is solved in a straightforward manner: individuals will support a norm whenever it gives them higher utility to do so.[7]

As noted, evolutionary game theory has the important substantive advantage of being a dynamic theory of behavior (Schotter 1981). Evolutionary game theory (Maynard Smith and Price 1973; Maynard Smith 1982) analyzes how populations of strategies change over time. The key assumption is that relatively successful strategies proliferate, while relatively unsuccessful ones dwindle. This core postulate refers *directly* to disequilibrium dynamics. Thus, unlike classical game theory, which has focused on behavior in equilibrium, evolutionary game theory can inherently deal with disequilibrium phenomenon.[8] In other words, evolutionary game theory can not only establish what is equilibrium behavior; it can also help us understand the dynamics that led to that behavior and forces that keep it stable. These properties preadapt it naturally to analyzing the emergence of norms.

Most of our results focus on explaining the emergence of *social norms* and, in particular, rules of behavior that prescribe third-party sanctions (in addition, of course, to dyadic, second-party punishments). Specifically, we use evolutionary game theory to show that strategies that encode social norms are necessary to stabilize behaviors in all nontrivial games (theorem 1); strategies lacking the social property of third-party enforcement cannot do the job. Moreover, for such strategies to be evolutionarily stable regardless of how quickly individuals adjust their behaviors, there must be an effective punishment for deviations from the norm (theorem 2). Stable strategies in such games must preserve the essence of their "social" nature throughout the game and throughout the relevant population (theorems 3, 4, and 5). Further, we show that some social norms, that is, the strategies

[6] Though our argument is close to Sugden's (1989; see also Sugen 1986), he does not focus on strategies that impose third-party sanctions. Few of the examples in his book are what we call "social strategies."

[7] For functionalist or efficiency-oriented analyses that take the microincentive problem seriously, see Ellickson (1991) and Calvert (1995).

[8] It is therefore important not to let the name, "evolutionary game theory," cause confusion: despite their semantic similarities, evolutionary game theory differs from classical game theory in several significant ways.

that encode them, are both evolutionarily stable *and* dysfunctional, in the sense of being Pareto inefficient (theorem 6). We also show, however, that functional (Pareto-efficient) social norms are *more* stable than dysfunctional ones (theorems 7 and 8). Thus, although a group can indeed dwell in a bad equilibrium underpinned by counterproductive social norms, long-run dynamics tend to drift toward more efficient ones. And qualitatively, there is a tendency for groups to drift toward an equilibrium regulated by social norms, instead of staying in equilibria where social norms have no part.

Substantively our results apply to a very wide spectrum of phenomena. If we were to agree with Ullmann-Margalit (1977) that norms are empirically relevant in three types of situations—prisoner's dilemma (PD) problems, coordination problems, and situations of inequality where the status quo benefits one party and hurts another—then most of our results cover all three types. Indeed, our earlier result (Bendor and Swistak 1998), as reinterpreted in this article (theorem 1), links stability with the necessity of social norms and covers literally *all* two-person games of any strategic interest whatsoever. Because, however, it is useful to motivate and illustrate the general argument via a specific example, we focus throughout on a particular game—the two-person PD, which is so well understood in the social sciences that little explication of the game is required here.[9] We hope that the example's specificity will not mislead readers into underestimating the generality of the analytical results that follow.

All of our results are based on evolutionary game theory. This approach diffused early and rapidly in the biological sciences and a bit later, though no less rapidly, in game theory, economics, and political science.[10] Sociology remains much less influenced by the rational choice paradigm in

[9] In the one-shot PD, each player has two choices, to cooperate or defect. Payoffs to each player in the one-shot game are as follows: R (when both cooperate), P (when both defect), T (to the defecting player when his opponent cooperated), and S (to the cooperating player when his opponent defected). The payoffs are assumed to satisfy $T > R > P > S$: defection yields a higher payoff regardless of the opponent's action, yet mutual defection is worse for both players than is mutual cooperation. Further, in analyses of the repeated game it is typically assumed that $2R > T + S$: mutual cooperation is better than alternately exploiting and being exploited.

[10] See, e.g., Hines (1987), Axelrod and Dion (1988), and Vincent and Brown (1988) for review articles in the biological sciences. For review articles focused on economics and political science, see, e.g., Friedman (1991), Selten (1991), Mailath (1992), and Samuelson (1993). There are also many books devoted in part or entirely to evolutionary game theory. These include Axelrod (1984), van Damme (1987), Hofbauer and Sigmund (1988), Binmore (1992), Bomze and Potscher (1989), Cressman (1992), Fudenberg and Levine (1998), Samuelson (1998), Vega-Redondo (1996), Weibull (1995), and Young (1998).

general and evolutionary game theory in particular.[11] Because it may still be relatively unfamiliar to many sociologists, we provide an introduction to game theory (classical and evolutionary). Next we lay out the central problem: the instability of dyadic strategies and what this implies about norms. Then, we show that stable nondyadic strategies do exist by introducing one strategy, conformity (or CNF), that uses an ancient social logic to stabilize itself. We then provide general results about the existence of stable strategies and their corresponding norms in all two-person games. In the following sections, we identify several essential properties of stable strategies and refute the strong functionalist thesis by showing that some stable norms are Pareto deficient. The functional thesis is partially redeemed later, when we show that more efficient norms are more robust than less efficient ones. In our two concluding sections we sketch some extensions, identify circumstances in which norm-encoding strategies do *not* enjoy evolutionary advantages, and give some brief applications of the results.

GAME THEORY AND EVOLUTIONARY ANALYSIS

There are two standard ways to clarify an abstract, mathematical argument. First, one could begin by constructing the simplest model and then extend it gradually to more complex ones. Second, instead of discussing the problem in general terms, one could illustrate it via a widely known and well-understood example. Since many sociologists are relatively unfamiliar with the language and tools of game theory we will use both methods to make the presentation as clear as possible.

Although evolutionary game theory is now a distinct mode of analysis, it grew out of classical, noncooperative game theory. Hence some very basic knowledge of the latter is useful for understanding the former. We begin, therefore, with some simple yet essential elements of noncooperative game theory, embedded in the substantive context of the "problem of cooperation" (as represented by the PD).[12] We then turn to evolutionary game theory.

[11] Rational choice still remains a rare mode of theorizing in sociology (Coleman and Fararo 1992; Hechter and Kanazawa 1997; Voss and Abraham1999) despite the influential early history of methodological individualism and rational choice (Homans 1950, 1961; Blau 1964), the attention garnered by Coleman's program (Coleman 1986, 1990*a*), and a growing number of conspicuous contributions from younger sociologists (e.g., Hechter 1987; Heckathorn 1988, 1990; Macy 1993; Macy and Skvoretz 1998). For some notable exceptions on the use of evolutionary game theory in sociology, see, e.g., Coleman (1990b), Kollock (1993), Lomborg (1996), and Raub and Weesie (1990).

[12] This explantion is based in part on Bendor and Swistak (1996).

Some Basic Game Theory

The problem of cooperation and its simplest model.—To better understand the nature of the general problem of norms, we begin our discussion with the simplest possible game-theoretic model (a one-shot, two-by-two game) of one of the most important problems of the social sciences—the problem of cooperation. Informally speaking, the problem of cooperation concerns explaining when rational actors will stably cooperate as they interact in situations involving conflict. Arguably, social interactions are fundamentally based on such exchanges. Many informal interactions, in small groups or networks for instance, turn on an exchange of favors, and such exchanges exhibit an important type of tension: egoistically, it is always best to obtain favors without returning them (doing a favor is costly), and the worst is to give and not get; yet both parties are worse off when favors are withheld than they would have been had favors been exchanged.

More specifically, consider two players, Row and Column. Row can either do Column a favor or refuse to do so; Column has similar options. Let's assume that if Row obtains a favor without returning one, then Row gets his best outcome, with a utility of $T = 5$. Assume, moreover, that the second-best outcome ($R = 3$) results from the mutual exchange of favors. If favors are withheld, Row gets the third-best outcome ($P = 1$). Finally, the worst outcome for Row is when Row helps Column but gets nothing in return ($S = 0$). For simplicity, we can further assume that the game is symmetric so that Column has identical utilities. With $T > R > P > S$ this is an instance and the simplest model of the famous PD game. Defection is a dominating strategy: it yields a higher payoff regardless of the other player's action. This makes mutual defection (no favors exchanged) the only equilibrium in the one-shot version of this game. Rational actors are, hence, bound to end up defecting with each other. The dreary outcome is that in equilibrium players get a payoff of ($P = 1$, $P = 1$) which is worse, for both, than the payoff to mutual cooperation—when favors are exchanged—of ($R = 3$, $R = 3$). Defection is individually rational but collectively suboptimal. (See fig. 1 for the choices and payoffs of the one-shot PD.)

A model with repeated interactions.—If the PD is played just once (and players know this and are rational) then the game's outcome is a Pareto-inferior set of payoffs; both players could have done better had they chosen to cooperate. Suppose, however, that individuals interact repeatedly, and they anticipate many future interactions with no definite termination date. (In a repeated version of the PD we will additionally assume that $R > (T + S)/2$, which means that the collectively optimal outcome results from mutual cooperation.) Formally, we can model this situation by assuming

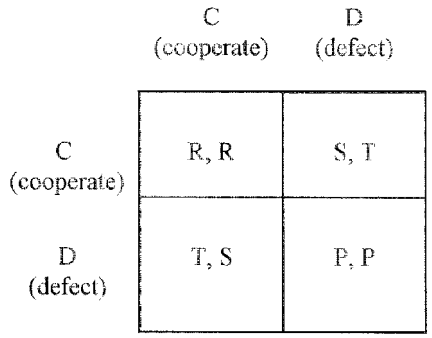

FIG. 1.—The prisoner's dilemma game: $T > R > P > S$ and $R > \frac{1}{2}(T + S)$ (row player's payoff is listed first in each cell).

that both players play an infinitely repeated[13] one-shot game with the same payoff matrix and the same probability of continuing the game in every period.[14] A (pure) *strategy* of player A in a game with B is defined as a complete plan for A; that is, a function which specifies A's move in any period k depending on B's moves toward A in periods 1 through $k - 1$.[15] Let $V(j,i)$ denote strategy j's expected payoff when it plays strategy i. The payoff to each strategy, $V(j,i)$ and $V(i,j)$, is then computed as the expected value of a string of payoffs resulting from the play between two strategies. For instance, if both j and i are "always defect" (ALL-D)—defect unconditionally in all periods—the payoff to each of them is computed as the following infinite sum:

$$V(\text{ALL} - \text{D}, \text{ALL} - \text{D}) = P + P\delta + P\delta^2 + P\delta^3 + \ldots = P\frac{1}{1-\delta}.$$

In general we can interpret δ as an effect of discounting future payoffs or a probability that the players will interact in the next period or a joint effect of uncertainty and discounting. Since all three interpretations of δ deal with the effects of the future, it is both natural and accurate to refer to a game with sufficiently high δ as *a game in which the future is sufficiently important*. We will use this less technical phrase throughout the article.

The reason we will focus on games for which the future is sufficiently

[13] See Rubinstein (1991) for why an infinite game can be used as a reasonable model of finite interactions.

[14] Note that with this assumption—a constant probability of continuing the game—the *expected* duration of the repeated game is finite and equals $1/(1-\delta)$.

[15] A mixed strategy in a repeated game is defined as any probability distribution on a set of pure strategies.

important is simple—the problem of cooperation is theoretically uninteresting in repeated games in which the future matters little. This is because an equilibrium in an iterated PD (IPD) with low values of δ is the same as the equilibrium in the one-shot PD—defection. The intuition behind the simple algebra of this result is simple too: if the future payoffs do not matter much, maximizing in a repeated game is really equivalent to maximizing in a current period of the game. Hence for low δ, the ALL-D strategy is the best response to all other strategies in the IPD. The emergence and the stability of cooperation become meaningful issues only in games where the future is sufficiently important. Hence from now on we will limit our analysis to this type of game.

To see what kinds of strategies can be supported in equilibrium in games where the future is sufficiently important consider, for instance, a strategy of tit for tat (TFT): cooperate in the first period of the game and thereafter in any period k do what one's partner did in $k - 1$. Consider two players, A and B. Suppose that player A is using TFT and player B, knowing A's strategy, is trying to maximize against it. If the future is sufficiently important (i.e., δ is high enough), then—given the retaliatory nature of TFT—it is best for B to cooperate with TFT in all periods. Defecting in any period would only lower the expected payoff.[16] Hence, if B were contemplating playing TFT against A, he would have an incentive to do so since using any other strategy would yield either a lower or, at best, the same expected payoff. If we now apply the same reasoning to the other player we see that neither player, given the other's strategy, has an incentive to change his own. In game theory this means that the pair of TFT strategies in the IPD with sufficiently important future form a *Nash equilibrium*.[17]

And so, the last paragraph provides two important observations. First, we have illustrated the notion of equilibrium in games—the so-called Nash equilibrium. And second, we have established that mutual cooperation can be sustained as a Nash equilibrium if the future is sufficiently important. In contrast to the dreary outcome of the one-shot game, this is clearly an optimistic conclusion.

The existence of a cooperative equilibrium does not mean, however,

[16] For instance, the strategy of ALL-D will score lower with TFT than would another TFT. To see this, note that $V(\text{ALL-D,TFT}) = T + P\delta + P\delta^2 + P\delta^3 + \ldots = T + P\delta/1 - \delta$. and $V(\text{TFT, TFT}) = R + R\delta + R\delta^2 + R\delta^3 + \ldots = R/1 - \delta$. If we take, e.g., a game with $T = 5, R = 3, P = 1, S = 0$ then $V(\text{TFT, TFT}) > V(\text{ALL-D, TFT})$ in all games where $\delta > 0.5$. In fact, any strategy that defects with TFT in any period of the game will do worse than a strategy that cooperates with TFT in all periods. For the proof that TFT is a best response (i.e., one that yields a maximal payoff) to TFT when the future is sufficiently important, see, e.g., Axelrod (1984).

[17] This term is used to honor John Nash, one of the main pioneers of game theory.

that a game will end with both parties cooperating. Whether it will or not depends on what other equilibria are possible in the game. The problem of identifying all equilibria in repeated games is the subject of a well-known result that goes by a cryptic name: *the folk theorems*.[18] While folk theorems apply to any repeated game—not just the IPD—we focus below on their interpretation and implications for the game of IPD—the case of our running example. Also, to avoid getting mired in more technical aspects, we will consider one very specific interpretation of the IPD folk theorems.

This interpretation concerns the "amount of cooperation" sustainable as a Nash equilibrium in the IPD. Suppose we measure the amount of cooperation in the IPD as the frequency of cooperative moves in the set of moves by both players. With this interpretation, the relevant folk theorems can be stated briefly: *any amount of cooperation can be sustained in equilibrium*. Hence, a pair of strategies in equilibrium can support anything between 0% cooperation (e.g., a pair of ALL-D strategies) to 100% cooperation (e.g., a pair of TFT strategies).[19] Thus, an equilibrium may contain permanent cooperation, or permanent defection, or any intermediate outcome between these two extremes.

What eventuates as equilibrium behavior depends, of course, on what player A thinks player B is playing and what B thinks A's strategy is. If A believes that B is playing ALL-D, A would have no incentive to do anything but to play ALL-D as well.[20] The same reasoning holds for B. Thus, if A thinks that B is playing ALL-D and B thinks that A is playing ALL-D, they will both keep defecting when playing each other. But this presents the following problem: What if A's beliefs about B or B's beliefs about A are wrong? Would their interaction correct this problem? Clearly not. As they both constantly defect with each other, the game's actual

[18] According to Aumann (1987, p. 20) one particular version of the folk theorems has been known since the late fifties. The name was invented because the theorem lacks clear authorship and game theorists knew about the result long before any published account of it appeared. Since the same effect of a proliferation of equilibria can be obtained under different assumptions, the plural form—folk theorems—is often used. By now there are many accounts of this result in the literature. The reader may want to consult any of the recent textbooks on game theory (e.g., Rasmusen [1989] provides a less technical account; Fudenberg and Tirole [1991] and Myerson [1991] present well-formalized, more mathematical accounts).

[19] Hence, using classical game theory it is impossible to rule out any outcome in the IPD that gives each player his/her "security level," i.e., the payoff that they can unilaterally obtain for themselves. (In the IPD, a player can ensure that s/he gets at least $P/1 - \delta$, simply by always defecting. Hence an outcome in which someone gets less than this value cannot be an equilibrium.)

[20] Note that playing ALL-D is the best response to always defect in all games, i.e., no matter how important future payoffs are.

play gives them no opportunity to disconfirm the belief that their opponent is playing ALL-D.

Consider, for instance, the following eventuality: suppose that A plays a suspicious version of TFT (i.e., STFT): defect in the first period and thereafter play what the opponent played in the previous period. Assume also that A thinks that B is using ALL-D. Suppose now that B does the same: he plays STFT and thinks that A plays ALL-D. Then both A and B will defect in every period of the game, making it impossible to disconfirm either player's hypothesis about his partner's strategy.[21] Consequently, these mistaken hypotheses will lock the players into the inefficient outcome of mutual defection—unless they have some additional information that could test their hypotheses about their opponent's strategy. *Such additional information can be provided by the social context of interactions.* Indeed, if interactions take place in a small group, for example, players can often observe what happens between others. And even if they do not directly observe the relevant actions, they will frequently learn about them through gossip. Information about these other interactions may often reveal that a partner's strategy is not at all what one might have thought it was. Thus a group may be a clearing house for information that allows individuals to learn and to adjust their behaviors in ways that would be impossible otherwise. This brings us, finally, to *evolutionary* game theory, which extends classical game theory by analyzing how individual behaviors will evolve as people learn about the behaviors of others in the group.

Evolutionary Game Theory

A game in a group and the evolution of behaviors.—Since folk theorems imply that it may be rational to play a great many strategies in the IPD, the issue of learning becomes fundamental in games with sufficiently important futures. A group supplies an important vehicle for learning about the strategies of others. Take, for instance, our example of two players, A and B, who kept defecting with each other while wrongly assuming that their opponent played an ALL-D strategy, whereas both A and B have, in fact, played a strategy of STFT. If A and B play against each other in this manner, they will never find out that their assumptions about the strategy of the other player are wrong; consequently, they will be unable to maximize against the opponent's true strategy of STFT.[22]

[21] This means that the strategies of both players are rationalizable (Bernheim 1984; Pearce 1984).

[22] Any strategy that generates mutual cooperation with suspicious TFT from period 2 on is (if the future is sufficiently important) a best response to that strategy.

This may not be so if there were another player in the ecology with whom A or B were able to establish cooperation. Suppose, for instance, that there is a player C who plays a strategy of tit for two tats (cooperate in the first two periods of the game and then defect if and only if the opponent defected in the past two consecutive periods; hereafter TF2T). Given the definitions of STFT and TF2T it is easy to see that C will induce mutual cooperation with both A and B from period 2 on. However, cooperation with C should lead A and B to revise their beliefs about each other. Consequently, A and B may want to change their strategies in a way that would allow them to cooperate with each other and, hence, maximize their payoffs in the games they play in this group. For instance, they may switch to playing TF2T, given this strategy's success in establishing cooperation with A and B. In general, having learned about others' strategies, *players may want to switch their own strategies to ones that were more successful*. This example depicts the core idea of an evolutionary game.

The evolutionary game.—As is common with important notions, the general idea of an evolutionary game is simple.[23] A game represents pairwise interactions between individuals in a group. The standard evolutionary model (Maynard Smith 1982) assumes that the interactions are unstructured: that is, every individual has the same constant probability of encountering any other member of the group.

Each pairwise interaction is modeled as a one-shot game with a specific payoff matrix. Since individuals interact repeatedly, each pairwise contest ends up being a repeated game between the two individuals. In the game-theoretic model of a pairwise repeated interaction, δ represents the continuation probability, that is, the probability that, having reached period t, the game will continue into period $t + 1$. A standard simplifying assumption is that the payoff matrix of the stage game (i.e., the one-shot game) stays the same across all iterations.

Once a game is specified, $V(i,j)$ denotes strategy i's expected payoff when it plays strategy j. Strategy i's overall score or *fitness*, denoted $V(i)$, is the sum of its pairwise payoffs. Consider, for instance, a group of three actors playing TFT, TF2T, and ALL-D. Each actor plays the other two in the group. For instance, TFT scores V(TFT, TF2T) in a game with TF2T and V(TFT, ALL-D) in a game with ALL-D. Thus V(TFT), the total score of TFT in this group, equals $V(\text{TFT}, \text{TF2T}) + \text{V(TFT, ALL-D)}$.

Evolutionary agents.—Evolutionary change involves actors adjusting their behaviors (strategies) over time using trial and error: they adopt strategies that seem to work and discard those that do not. The model is

[23] For a more detailed introduction to evolutionary analysis, see, e.g., Weibull (1995).

dynamic. While the setup is essentially game theoretic, actors are not assumed to be the spooky *homo ludens*—"well-informed mathematical prodigies capable of costlessly performing calculations of enormous complexity at the drop of a hat [who assume] that [their] fellows are prodigies like [themselves], [and continue] to hold [that] belief whatever evidence to the contrary [they] may observe" (Binmore and Samuelson 1994, pp. 45–46). Individuals are rational agents but boundedly so—their goal is to maximize, but they are capable of making mistakes.[24]

How, then, will these boundedly rational agents adjust their behaviors? The answer to this question turns, naturally, on what they can observe. If the only observable conditions of the group are the performance and the relative frequencies of the strategies used by the actors, then the $V(j_k)$'s and p_k's would be the only factors that could affect replication of behaviors. In other words a dynamic process that governs replication can only be a function of strategies' fitnesses and their frequencies.

Evolutionary dynamics.—So far the evolutionary model describes what happens in a group within a block of time which corresponds to a life span of players' strategies or, speaking informally, their norms. We will refer to this timespan as a generation. Within a generation players learn about each others' norms. Across generations, they adjust their behavior given what they have learned about the group so far. The essence of evolutionary dynamics is simple: the more fit a strategy is in the current generation, the faster it increases. In other words, an evolutionary process is a dynamic that is increasing in fitness. We call this the *fundamental evolutionary postulate*; any dynamic with this property will be called *an evolutionary dynamic*. Note that this axiom is about how strategy frequencies change over time, it is not an equilibrium condition. As we have emphasized earlier, *disequilbrium dynamics are central for evolutionary game theory*.

The mechanisms driving change of behaviors vary with the domain of application. In biology it is typically assumed that the mechanism is genetic (Dawkins 1989). In the social sciences, evolutionary game theory postulates that behavioral processes, such as learning through, for example, imitation or socialization (Axelrod 1984; Gale, Binmore, and Samuelson 1995; Boyd and Richerson 1985; Cabrales 1993), are the driving

[24] For a detailed discussion of the connection between the ideas of bounded rationality (Simon 1957; March and Simon 1958) and evolutionary game theory, see Mailath (1992).

forces. Thus individuals learn to discard behaviors (strategies) that yield low payoffs and switch to strategies with high payoffs.[25]

Equilibria.—The "dual" of dynamics is stability, for the standard notion of evolutionary stability is tightly connected to a dynamical analysis. Under what conditions should we consider a strategy to be evolutionarily stable? Consider a group where all individuals play the same strategy; let's call it a "native" strategy. Because everyone is using the same strategy, everyone gets the same payoff; hence the ecology is in equilibrium. Now perturb it slightly by allowing a few new behaviors to invade the group. (These new strategies are often called "mutants," reflecting the biological roots of evolutionary game theory.) An evolutionarily stable strategy is one that, once it is sufficiently common in a group, can resist this kind of small perturbation or invasion. More precisely, strategy i is evolutionarily stable if there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$ the population playing the native strategy i can resist any ϵ-size invasion of mutants.

Intuitively, "resist an invasion" may be understood in two ways: stronger, if after the invasion the invaders decrease in frequency under the evolutionary dynamic, and weaker, if they do not increase. Since in iterated games the best the native strategy can do is prevent the mutant from spreading, weak stability *is the only type of stability attainable* in repeated interactions (Selten 1983; van Damme 1987). More specifically, we will call a strategy *weakly stable* (Bomze and van Damme 1992) if it

[25] It is sometimes said that assuming that players know each others' strategies is unrealistic and hence conclusions based on this assumption are of dubious value. We would like to briefly address this important concern. The root of the problem turns on the assumption which allows *any* theoretically possible strategies into the analysis. Clearly, some strategies are much too complex to be considered as reasonable models of human behavior. This, however, is not true of all strategies; many are exceedingly simple. TFT, e.g., is defined by two very simple properties, reciprocity and retaliation, and if a player is using TFT it is quite realistic to assume that her partner knows what she is doing. And in general, if strategies are simple there is nothing unrealistic in assuming that players know each others' strategies. Indeed, the common knowledge of norms is an essential part of a group's culture. Thus, it is natural to think about a group in which it is common knowledge that its members are, for instance, "reciprocating" and "retaliatory" even if one of these properties remains latent (e.g., players who always cooperate with each other would never retaliate). That is, it is quite natural in such cases to assume that players may know what to expect of others even if they have never interacted before. Hence if we confine our attention to a set of cognitively feasible, simple strategies we find nothing unrealistic about the assumption that players know each others' strategies. Moreover, it is important to point out that our analysis will produce *equilibria that can be supported by such "simple" strategies*. Thus allowing all theoretically possible strategies into the analysis ultimately has no bearing on the empirical soundness of the result. Another way of dealing with this problem is to assume that players do not use strategies but rather simple behavioral rules of thumb which are not "complete plans of the game." We develop such a theory elsewhere (Bendor and Swistak 2000).

does not decrease in frequency in any group (with a finite number of strategies) where its frequency is sufficiently high.[26] In this case certain mutants may have the same fitness as the native and hence may remain in the population indefinitely.

Finally, it is important to emphasize that the fundamental evolutionary postulate does not define a single process; instead, it defines a *class* of processes. Thus a native strategy may be stable under one dynamic or a class of dynamics, yet unstable under another. Hence sweeping claims that "strategy x is evolutionarily stable" always have to be predicated on the class of dynamics under which x is, in fact, stable.

It is also important to note that the form of dynamics is directly related to certain fundamental social characteristics of people in the group. The rate at which an individual adjusts his behavior can be thought of as a measure of the inertia of the norms he holds. This inertia is affected by two principal determinants of human behavior—the sociological factor of group conformity and the economic factor of payoffs. Thus a pure *Homo sociologicus* would completely disregard payoff considerations and would only react to the conformity factor, or the proportion of the group that shares his norm—the more people share his norm, the higher his utility for the norm. On the other extreme there is *Homo economicus*, a player who reacts only to payoff considerations—the higher the payoff to a norm, the higher his utility for the norm. (See Bendor and Swistak [1997] for a more detailed discussion and a model of these effects.) In between the two ideal-types of *Homo sociologicus* and *Homo economicus* lies a large spectrum of behaviors where players' incentives are partly affected by conformity and partly by payoffs. A specific form of a dynamic process is a function of the specific set of incentives held by people in the group. The dynamic will be different if all actors are the pure *Homo sociologicus* type, or if they all are *Homo ecomomicus*, or if they all have mixed incentives, or if a group consists of some *Homo sociologicus*, some *Homo economicus* and some mixed types, and so on.

Robust equilibria.—There is, however, one happy circumstance in which the composition of individuals' incentives in a group does not matter. Suppose strategy i is *unbeatable*: it gets the highest payoff once it is sufficiently common. More precisely, consider any group in which strategy i is played with frequency p_i. We refer to i as unbeatable if there is a $p^* < 1$ such that for any group where $p_i > p^*$, $V(i) \geq V(j)$ for any strategy j. Clearly, when i is unbeatable it must also be (weakly) stable under *all* evolutionary processes. And it is easy to see that the reverse claim holds as well: if a sufficiently frequent strategy is weakly stable

[26] Others have referred to a weakly stable strategy as semistable (Selten 1983), neutrally stable (Sobel 1993), or neutral evolutionarily stable strategy (Warneryd 1993).

under all evolutionary process, it must be unbeatable. Hence, the existence of an unbeatable strategy confers two great methodological boons: first, it is unnecessary to specify a replication process in order to study the qualitative properties of stability; second, the stability of an unbeatable strategy is unaffected by changes in process specification. No matter what specific dynamic holds in a given domain, once an unbeatable strategy has become sufficiently common, it will not decrease.

There is, however, much more to the idea of seeking stability under all processes than a simple consideration of methodological robustness; this type of stability may be necessary for an equilibrium to be descriptively meaningful. There are at least two reasons for this claim. First, because how strategies replicate reflects the way people adjust their behavior and because different individuals adjust in different ways and may change the way they adjust, an equilibrium that cannot be sustained under all processes may be highly unstable. Second, a mix of learning and imitation, in the face of both social and economic incentives, may affect different people in different ways that are often complex and hard to pin down. Hence, assuming that the process is described by a specific type of dynamics seems infeasible. Empirically it may be impossible to discern precisely how strategies replicate or what equations approximate the dynamics. In conclusion then, equilibrium states that are empirically meaningful may have to be states that are stable under all processes. We will refer to strategies that are unbeatable, or equivalently (weakly) stable under all processes, as *uniformly stable*. It is this type of stability that we will seek to establish below. It is interesting that the answer to the question, Why are there norms? is tightly linked to the concept of uniform stability.

THE ESSENCE OF THE PROBLEM AND THE CONCEPT OF A SOCIAL NORM

Norms are ordinarily thought of as rules specifying actions regarded by a group as proper (prescriptive norms) or improper (proscriptive norms). It is typical, hence, to understand norms as codes of behavior regardless of whether this behavior involves merely two individuals[27] or more.

While we have no argument with such a general conception of norms, we would not find the emergence of some rules belonging to this class particularly puzzling. Take the example of PD. It is not surprising that two actors locked in a long-term relationship may choose to cooperate in this game. It is quite clear that mutual cooperation can be supported in

[27] It is common to talk about a norm of "reciprocity" or a norm of "retaliation" (Axelrod 1984; Bicchieri 1993; Ullmann-Margalit 1977), a "TFT" norm of behavior, etc.

equilibrium by, for instance, a pair of TFT strategies: if one player believes his partner is playing TFT, he cannot improve his payoff by playing anything but TFT himself (Axelrod and Hamilton 1981; Axelrod 1981, 1984). This then explains the norm of "reciprocity" and "retaliation" both of which are properties of TFT. Since this explanation is pretty straightforward, the associated norms are not particularly surprising. It is norms that cannot be explained this way that are puzzling.

Hence, we see the essence of the puzzle of norms as turning on rules that make us reward and punish other players not for what they did to us but for what they did to other members of the group. In general, we will say that a behavioral rule followed by i is a *social norm* if i conditions his behavior toward some j not only on what has transpired between i and j but also on other interactions in the group to which i and j belong. A norm that says that a friend of your friend should be your friend or one that says that a foe of your foe should be your friend are examples of such rules. By the very nature of the interactions involved these types of rules cannot be reduced to pairwise interactions; they are intrinsically group phenomena. Thus social norms are social sine qua non. These are the types of norms that we want to explain.

Consider a very specific instance of the general problem. Take, for example, three people, A, B, and C, and assume that they all interact with one another (pairwise) in an IPD. Suppose that initially all actors cooperate with others in all interactions: A cooperates with B, B cooperates with C, and A cooperates with C. Assume, however, that at some point A starts defecting toward B whereas B loyally continues to cooperate with A. Assume, moreover, that as B is being exploited by A, cooperation between A and C and between B and C continues. C's problem is now clear: should C punish A (by defecting toward A) for what A does to B or not? The norm that says that foes of your friends ought to be your foes requires that C defect toward A. But if C punishes A for what A does to B, C may jeopardize the beneficial cooperative exchange she has with A. By complying with the social norm, C would be decreasing her expected payoff—it may not be rational for C to do so. Indeed some have claimed (Elster 1989*a*) that the very essence of such norms is that they cannot be explained by rational considerations. It is precisely this type of behavior—one that involves third party sanctions and rewards—that is the focus of this article. Thus when we pose the question, Why are there norms? it is norms in this specific sense of "social norms" that we have in mind. To remind the reader about this qualified meaning we will sometimes say social norms instead of norms. Whether we say it or not, however, the article focuses on this specific sense of norm.

THE PROBLEM: FREE RIDING ON PUNISHMENT AND DEFENSE

If we look more closely at the nature of the problem that faces the three actors in the example above, we would discover that complying with the norm creates a standard public good problem, also known as the second-order collective good problem (e.g., Axelrod 1986; Coleman 1990*a*; Flache and Macy 1996; Hardin 1995; Heckathorn 1990; Oliver 1980). Everyone is better off if the norm is obeyed, yet no one has an incentive to comply. The public good, in this case, is a code of behavior—a norm.

The essence of the problem is best explained by an example (Boyd and Lorberbaum 1987) which extends the story of players A, B, and C above. Suppose people play each other in pairwise games of the iterated prisoner's dilemma (IPD). Initially everyone plays TFT. Because everyone is using a *nice* strategy—one that never defects first (Axelrod 1984, p. 33)—each player cooperates with everyone else in all periods. Therefore TFT's provocability, its readiness to punish defection, is never tapped. Because a latent property may decay, some TFT strategies may mutate into less provocable ones, say TF2T (cooperate in periods 1 and 2 and then defect if and only if your opponent defected in the previous two periods.) Now most people are playing TFT; a few use TF2T. Because TF2T is, like TFT, nice, everyone continues to cooperate with all partners; hence the introduction of this new strategy is, as yet, unobservable. (Henceforth, invaders that are observationally indistinguishable from the native strategy will be called *neutral* mutants.)

Now a behaviorally distinct strategy, STFT, appears. The nonnice cousin of TFT, STFT defects in period 1 and thereafter reciprocates its partner's previous move. Because STFT is not nice, it is a genuine threat to the community's code of constant cooperation. TFT responds to this deviation by punishing STFT in period 2. Unfortunately, this triggers a vendetta, with STFT defecting in odd periods and TFT in even ones. Thus, though TFT's punishment is effective—if most people use TFT then STFT will perform the worst in the ecology—it also hurts TFT: a vendetta is not the best reply to STFT, if the future matters enough. (More precisely, TFT's behavior is not a best response if δ, the continuation probability, is above a critical threshold and $2R > T + S$.) Under these conditions, the best response to STFT is to ignore its first defection and to cooperate thereafter—exactly how TF2T behaves.

Thus in this case, the burden of defending a "nice" code falls on players using TFT; those using TF2T free ride on this enforcement. The result of this free riding is that, if δ exceeds the critical threshold, TF2T obtains the highest payoff: $V(\text{TF2T}) > V(\text{TFT}) > V(\text{STFT})$.

The dynamic implications of this outcome depend, naturally, on the specifics of strategy replication, on how players adapt. Assume, for in-

stance, that players in the current generation are motivated only by payoff considerations and they adjust by switching to the strategy that did best in the previous generation, hence imitating the winner. If this is how players adjust their strategies, in the next generation everyone in the community will play TF2T. And so the simple dynamic of imitating the winner destabilizes TFT, as its willingness to punish miscreants proves costly. Thus this famous strategy *would not replicate itself* in this selection environment.

Further, this example generalizes enormously, as has been shown by Bendor and Swistak (1998) in the following theorem 1. This result contains two terms that need defining. First, a one-shot game is *trivial* if it has an action that yields the game's maximal payoff no matter what the other player does. Second, a *dyadic strategy* i is one that bases its behavior toward another strategy j only on the history of play between i and j. Information about the history between j and a third strategy k is never used by i in its conduct toward j. (As a convenient shorthand, we will sometimes ignore the difference between a strategy and a player using that strategy.) If strategy i is not dyadic, then its conduct toward j will at least sometimes be based on what has happened outside the (i,j) relationship. Hence we will call these *social* strategies.[28]

THEOREM 1 (Bendor and Swistak 1998).—*In any repeated nontrivial game in which the future is sufficiently important no pure dyadic strategy is uniformly stable*.

Because trivial games have no strategic interdependence whatsoever, they are trivial in every respect. Practically, then, the theorem covers every kind of two-person interaction that might be of substantive interest to social scientists. Thus, in all that follows, we restrict attention to nontrivial games. To avoid tedious repetition, statements of results will omit this restriction. In the corollary that follows, for instance, the phrase "for any repeated game" should be read "for any repeated nontrivial game."

To understand why social norms evolve, it is useful to state a corollary of theorem 1.

COROLLARY.— *If a pure strategy is uniformly stable in a repeated game in which the future is sufficiently important, then it must be a social strategy.*

Logically speaking this corollary merely restates the theorem, but it

[28] The connection between the formally defined idea of social strategies and the less formal notion of a social norm should be clear: since a strategy is a formal, game-theoretic model of a norm, *social strategies represent social norms*. Dyadic strategies, which by definition are not based on third-party play, cannot represent social norms. (We will sometimes take advantage of this correspondence by slipping back and forth between "social norms" and "social strategies" in the text. The distinction is maintained rigorously in the appendix.)

emphasizes a different set of strategies: those that encode social norms. Thus the corollary to theorem 1 gives part of the solution to the question, Why are there (social) norms? This part of the answer is simple: if there are any strategies that are stable in the face of all evolutionary dynamics, including harsh processes driven by extreme economic incentives as when everyone imitates the winner, then *they must encode some type of social norm*.

WHY MUST FOES OF YOUR FRIENDS AND FRIENDS OF YOUR FOES BE YOUR FOES? A THEORY BEHIND THE FOLK WISDOM

By showing that social norms are necessary for stability, we have corroborated half of Sugden's (1989, p. 97) conjecture that norms "have evolved because they are more successful at replicating themselves than other patterns [of behavior]." To corroborate the rest of his conjecture, we must show now that in these same settings, there are social norms that *do* replicate reliably. We begin with an example and generalize it below.

Let us return to the previous example of Boyd and Lorberbaum to see what kind of norm or strategy might work. For simplicity we replace TFT by the unforgiving grim trigger (GT); the mutants continue to be TF2T and STFT.[29] As in the original example, the native loses to TF2T because the latter maintains a cooperative relationship with both of the other strategies, whereas GT and STFT are locked in mutual defection from period 3 on. If, however, the native would punish TF2T for tolerating STFT's uncalled-for defection in period 1, then the native would outperform both mutants. A norm that would require the native to do that is easy to imagine. Consider a new native, GT′, which modifies the old one in the following way: This native regards any strategy that is never the first to defect in a group as *socially nice* (all nice strategies are socially nice but, as we shall see, some socially nice strategies are not nice). GT′ is itself socially nice, and it defects forever against any strategy that either is not socially nice or that fails to punish one that is not socially nice. Thus GT′ defects from period 2 onward against STFT, and it punishes TF2T from period 3 onward because TF2T failed to sanction STFT, which is not socially nice. It is easy to show that, given any sufficiently small invasion of any two distinct mutants, GT′ will have the highest fitness. (If both invaders are socially nice, then all three strategies are observationally indistinguishable and so will be equally fit.) The new

[29] GT is defined as cooperate in period 1, and cooperate in period $k > 1$ if and only if the partner has cooperated in all previous periods.

native, by being socially nice but not nice—it defects first against TF2T—has the necessary flexibility to outperform TF2T.

But this is not a general solution. GT′ can still be beaten by an arbitrarily small invasion. Ironically, the cause of its downfall is its "ancestor," GT. Consider a small invasion by TF2T, STFT, and GT. GT behaves just as GT′ requires: it always cooperates with the native, and it immediately punishes the strategy that is not socially nice, STFT. Therefore the native and GT cooperate forever. However, because GT is nice, whereas GT′ is only socially nice, the former cooperates constantly with TF2T, whereas the latter sinks into continual mutual defection with TF2T from period 4 onward. Hence if the future is sufficiently important, GT will outscore its more sophisticated kin.

Consequently, in order to outdo GT as well, the native must include a higher order rule, punishing GT for failing to punish TF2T's failure to punish STFT. Call this new native GT″. But the reader has probably guessed that GT″ is itself vulnerable to an invasion that includes the previous three mutants plus the ancestor of GT″—GT′. By the preceding logic, GT′, by punishing TF2T and STFT but maintaining good relations with GT, will score higher than GT″ no matter how small the invasion. But this will destabilize the native strategy, and the reasoning enters another loop.

It is evident, therefore, that for this kind of strategy to be uniformly stable, it must have a *metanorm* structure, punishing all *n*th-order deviations from its code. Player 1, who failed to punish player 2, who failed to punish player 3, and so on, must also be sanctioned. It is of course logically possible to construct such codes. But doing so may not only seem a rather daunting violation of the spirit of evolutionary game theory, which is informed by a strong dose of bounded rationality (Mailath 1992); it also looks completely utopian from the perspective of *any* model of human cognition and behavior.

HOW FRIENDS BECOME FOES—A NORM THAT SOLVES THE PROBLEM OF COMPLEXITY

The answer to the perplexing information problem that would require constant monitoring of the entire network of interactions and perpetual analysis of its complex metanorm structure lies in the way *Homo sociologicus* organizes information and acts upon it. The answer is a specific social norm. This norm corresponds to a very well known social mechanism.

The essence of the solution is exhibited by the following new strategy that we shall call conformity (CNF). This strategy is based on a binary

categorization of all players in a group (Heider 1958; Lévi-Strauss 1974). In every period every player is considered either a friend or a foe. In the beginning, before play starts, everyone is deemed a friend. If someone is a friend in period t, they remain a friend in $t + 1$ if and only if they cooperated with all other friends and defected against all foes in period t. Anyone who ever violates these rules becomes a foe thereafter. The following result is then easily established.

PROPOSITION 1.—*If the future is sufficiently important, CNF is uniformly stable in the IPD.*

The proof is straightforward. For any invasion, either a mutant j behaves exactly the same as CNF does or it behaves differently with at least one partner in at least one period. If the former, then j must obtain the same payoff as the native. If the latter, then j must have either defected toward a friend or cooperated with a foe, at some date t. In either case, CNF will punish it thereafter. Given CNF's response, the best that j can do is to defect against CNF from t onward. For the standard reasons, when the future is sufficiently important it is better to cooperate always with CNF than it is to cooperate in a finite number of periods, gain a one-time advantage of the temptation payoff T, and then get stuck in mutual defection thereafter. Hence CNF does better with itself than does any such j, which for sufficiently small invasions ensures that $V(\text{CNF}) > V(j)$. Therefore CNF is uniformly stable.

CNF has several appealing behavioral and cognitive properties that correspond to some well-known social mechanisms. It embodies, for instance, an ancient political logic that is very intuitive: the friend of my foe is my foe, the foe of my friend is my foe, and the friend of my friend is my friend.[30] (One might view CNF as trying to impose Heider's balance principles [1958, pp. 200–207], or perhaps more precisely, Davis's [1967] cluster principles, on its environment.) Thus at any given time the strategy partitions the entire community into two camps: allies and enemies. From a cognitive perspective, CNF, although it embodies a complete metanorm structure, is simple enough to be represented as a strategy of little more complexity than tit for tat: to know a partner's standing in the community tomorrow one only need keep track of that partner's status today and, if the partner is currently a friend, his current actions. Social mechanisms like gossip greatly reduce the burden of the second requirement—news about defections among friends spread amazingly quickly in some very large and complex networks. (Due to its unforgiving nature, if the partner

[30] Note, however, that CNF does *not* impose the fourth relation of this logic, i.e., it does not mandate that foes of foes are friends. This is fortunate for the empirical status of the theory since there is more evidence for the first three relations than there is for the claim that foes of foes must be friends (e.g., Taylor 1970, p. 204).

is a foe today, CNF does not need to observe that player's current actions in order to know his status tomorrow.) Thus the cognitive demands that solve the seemingly hopeless problem of complexity turn out to be surprisingly weak, well within the demands of bounded rationality.

"COVENANTS WITHOUT THE SWORD ARE BUT WORDS"

CNF in the context of the PD has a number of appealing properties and provides several invaluable insights. It clearly illustrates how a social norm can work and proves that a uniformly stable strategy exists in at least one game (the IPD). Of course, the problem is much more general, going far beyond the logic of the IPD. Fortunately, it turns out that one can use CNF's basic logic to establish that uniformly stable social strategies exist in several large classes of games.

CNF has three key elements: (*a*) a friend-foe categorization of all partners which is applied recursively; (*b*) a socially desired action and a punishment action in the one-shot game; and (*c*) a rule of playing the desired action toward friends and punishing foes. This logic carries over, with a few modifications, to almost all types of two-person games. Consider games that are symmetric.[31] In each stage game there is a finite set of actions $a_1, \ldots, a_m$. For notation, $v(a_t,a_s)$ denotes the payoff obtained by playing action a_t in the stage game against action a_s, and $\max_t v(a_t,a_s)$ is the payoff of the best response to a_s. The subset of symmetric games central to our next result is defined by two of Hobbes's essential elements: first, covenants are possible; second, there are swords for sanctioning people who break covenants. To see how these two elements are related in symmetric games, focus on those payoffs produced when the players take identical actions (payoffs on the main diagonal of the payoff matrix), and consider an action which yields the highest payoff on this diagonal. Since this may reasonably be deemed a "cooperative" outcome, let us denote this action as a_c and its payoff—$v(a_c,a_c)$—as R, to remind us of the PD notation. Now examine the subset of symmetric games with the following property: there exists a punishment action, a_d, such that R exceeds the best response to the punishment, $\max_t v(a_t,a_d)$. We will refer to a game with such a punishment action as *a game of enforceable cooperation*. The next result shows that this property—a sword that can enforce

[31] A one-shot game is symmetric if its payoff matrix is symmetric: the players have the same action sets and reversing the actions reverses the payoffs. Further, it is typically assumed that the game is *informationally* symmetric as well: players have no information that would allow them to distinguish each other as being in different positions, i.e., one as being the Row player and the other, Column.

a covenant—is decisive for the existence of pure uniformly stable strategies and the norms they encode.

THEOREM 2.—*Pure uniformly stable strategies exist in a repeated symmetric game with sufficiently important future if and only if the stage game is one of enforceable cooperation.*

It is easy to see why sufficiency holds. If the crucial payoff condition holds, then one can construct a native exactly on the lines of CNF: cooperate (play a_c) toward friends and punish foes with a_d. If anyone ever breaks these rules, they are a foe from the next period onward. This conduct implies that if a mutant ever acts differently with any partner than the native does, the native will punish it ever afterward, ensuring that the native does better with other natives than the mutant does with the native (when the future is sufficiently important). Hence the native's overall fitness must exceed the mutant's, if the natives are sufficiently numerous. Proving necessity is more troublesome; see the appendix for the proof.

Many repeated symmetric games studied by social scientists are covered by the above result. For example, in a game of iterated chicken, a uniformly stable social norm is to always be conciliatory with friends and always be aggressive toward foes. Since the payoff from mutual conciliation exceeds the value of the best response to aggression, a uniformly stable pure social strategy exists in this game.[32]

For a symmetric game that is not a game of enforceable cooperation, consider a common problem in institutions, the division of labor. The game illustrated in figure 2 presumes that the organizational goal will be reached if and only if the players succesfully divide the labor and thus complete the task. Conflict arises because one job confers higher status than the other. Thus, although each player prefers performing the low-status task to not attaining the goal, each prefers the structure in which he or she is top dog.

Although this game is symmetric in terms of actions and payoffs, the "name of the game" is specialization. Hence any pure native gets a low payoff when it plays itself. It is therefore easily invaded by a pure mutant that always takes the action that complements the native's choice.[33]

[32] There is a natural interpretation to punishing doves or "softies" in chicken: if aggressors are true invaders and softies are weak-kneed allies who refuse to fight them, then punishing doves is a case of moralistic aggression (Trivers 1971; Frank 1988).

[33] The division-of-labor game illustrates contexts which we might expect to become *informationally asymmetric*. As Sugden (1986) suggests, even though the game is symmetric substantively—the payoff matrix of the one-shot game is symmetric—players might recognize seemingly superficial differences in their positions. It is not hard to show that if the game is informationally asymmetric, then pure uniformly stable strategies do exist in games such as those in fig. 2. The case of asymmetric games is also

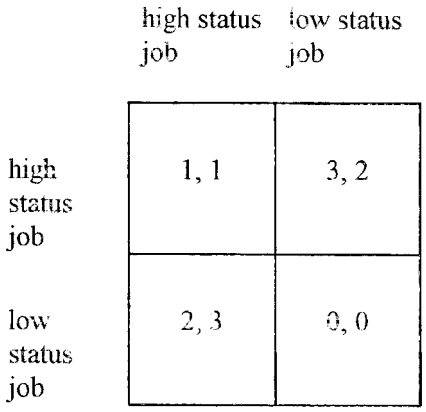

FIG. 2.—The division-of-labor game

GENERALIZING CNF'S PROPERTIES: THE ESSENTIAL FEATURES OF UNIFORMLY STABLE NORMS

Having shown that uniformly stable strategies and their corresponding norms exist in a wide variety of games, it is time to analyze their essential properties. A natural place to start is with our old standby CNF, which has proven so useful in providing insights into the nature of uniformly stable strategies.

The proof that CNF is uniformly stable does not imply that all uniformly stable strategies must be similar to CNF. In particular, it is important to point out that *a uniformly stable norm need not be as relentlessly unforgiving as CNF*. A strategy that is uniformly stable does not have to punish deviants forever for a single violation of a group norm. If the punishment is enough to wipe out the violation's gain and the deviant has returned to a proper behavior, a uniformly stable strategy may return to mutual cooperation or some other pattern of behavior required by the group norm.

Naturally, how much punishment suffices—to wipe out a violation's gain and to stabilize the native strategy—depends on the specific payoffs of the stage game. For some stage games the payoffs permit the corresponding rules of punishment to be particularly simple. For instance, in

---

easy to solve. The following counterpart of theorem 2 holds for asymmetric games: THEOREM 2′.—*Pure uniformly stable strategies exist in a repeated asymmetric game with sufficiently important future if and only if punishment is effective, i.e.,* $v(a_c,b_c) + v(b_c,a_c) > \max_t v(a_t,b_d) + \max_t v(b_t,a_d)$ *in the stage game.* In this formulation $a_1, \ldots, a_m$ denote Row's actions in the stage game; $b_1, \ldots, b_n$ are Column's (m need not equal n, but we assume both are finite); a_c and b_c are socially desirable actions, and a_d and b_d are punishment actions.

any IPD one can easily construct a uniformly stable strategy i that forgives, and does so in a rather natural way. Here is how i works. It uses the standard friend-foe categorization and cooperates with friends. If j (say) deviates, however, by defecting with i's friends or cooperating with i's foes, or if i's foes cooperate with j, then i should punish j until j's gains from violating the norm have been wiped out: as it turns out, in the IPD it suffices for j to submit to punishment—cooperate while being punished by i—as many times as it had violated the code. (Thus if j defected against three friends of i, then it would have to submit to punishment three times. In the IPD this would mean receiving three payoffs of S, to offset its prior ill-gotten gains of three T's.) Once j has paid his debt, however, the transgressions are forgiven and mutual cooperation can be restored. Hence we conclude that the unforgiving quality of CNF is not an essential property: a strategy does not have to be unforgiving to be uniformly stable.[34]

Which, then, of CNF's properties are really essential? What features *must* a strategy have if it is to be uniformly stable in repeated symmetric games? The above remark about other, more forgiving norms suggests that few of CNF's *dyadic* properties are essential. (This point will be reinforced by theorem 6 below, which shows that uniformly stable norms need not be "cooperative.") We already know, however, that to be uniformly stable a pure strategy must be social: it must attend to at least *some* third-party play in at least some periods. CNF, for instance, *never stops* monitoring third-party interactions and does not exclude any player from monitoring. (Even foes must be monitored, in their interactions with friends.) The next two results show that these two aspects of CNF—its temporal endurance and the social breadth of its monitoring—are not extreme; they are essential.

We formalize the definition of these monitoring features (time and social space) as follows. First, we will call a strategy *perpetually* social if it does not become dyadic after a finite number of periods. Second, we will call strategy i *comprehensively social* if there is no strategy j and period t in the game such that in all periods following t, i's actions toward other players in the group are independent of their interactions with j. Note that if this condition did not hold, that is, i were not comprehensively social, then following some period t i's actions toward a certain j would have been independent of j's interactions with other players. But this is equivalent to saying that, following period t, i would not need to monitor

[34] That being unforgiving is inessential is good news, because this property causes problems if partners' actions are imperfectly observed, as we will see later in this article. Fortunately, there are other uniformly stable social norms that are both forgiving and which sustain efficient outcomes as the stable state.

j's interactions at all. (The proofs of the following results are in the appendix.)

THEOREM 3.—*In any repeated game of enforceable cooperation in which the future is sufficiently important, all uniformly stable strategies must be perpetually social.*

This temporal property is required because otherwise a native could be invaded by a neutral mutant that patiently waited for the native to stop monitoring indirect pairs. Hence eternal vigilance is the price of stability. The next result shows that the norm embodied in a uniformly stable strategy must brook no exceptions in the community.

THEOREM 4.—*In any repeated game of enforceable cooperation in which the future is sufficiently important, all uniformly stable strategies must be comprehensively social.*[35]

Social comprehensiveness is required because if A_1 stops monitoring interactions of A_k then player A_2 may behave identically as A_1 toward everyone in the group except A_k with whom he may score higher than A_1. If this happens, A_1's strategy i would lose to the strategy of A_2, rendering i not uniformly stable. As noted, CNF monitors all of its partners, indefinitely. Theorem 4 shows that the same is required of every uniformly stable norm: it must be comprehensively social.

We now turn to examining the internal logic of uniformly stable norms. Again, CNF provides a clue. Recall that any CNF-like strategy that lacked a *complete* metanorm structure—failed to punish some nth-order deviation—could be invaded. The same logic applies to all uniformly stable norms.

THEOREM 5.—*In any repeated game of enforceable cooperation in which the future is sufficiently important, every uniformly stable strategy must have a complete metanorm structure.*

The proof rests on the existence of mutants that reply to punishment with punishment (i.e., by playing a_d). When encountering such invaders, the native-punisher gets $v(a_d,a_d)$, which is less than $v(a_c,a_c)$, since it is a game of enforceable cooperation. So punishing such mutants is costly. Thus there is a potential advantage to free riding on enforcement: If the native, i, does not punish j's failure to punish a lower-order deviation by

[35] In our conceptual framework, strategies (equated here with players) can distinguish amongst each other only on the basis of their actions; players lack ascriptive identities. Hence if two players have behaved identically toward everyone in the group then other strategies must treat them identically since they cannot tell them apart. A natural extension of this framework would endow strategies (players) with such ascriptively based information, so that they could behave differently toward people who have behaved identically (Bendor and Swistak 1999). It is worth noting that theorem 4 still holds in this more general formulation.

k, then i can be beaten by j, since the latter avoids paying the cost of enforcing the code.[36]

It is important to emphasize here that although requiring a complete metanorm code sounds rather daunting, it is actually easy to implement such rules: one simply codes all members of the group as being in either good or bad standing and treats them accordingly, thereafter applying a straightforward recursive logic to update an individual's social standing. The recursive updating enables the metanorm structure to be complete yet simple enough to be implemented even by very boundedly rational agents. Certainly CNF, a paradigmatic example of a code with a complete metanorm structure, is very simple.

Indeed, humans are capable of creating and using far more complex codes; CNF, which merely illustrates how to construct a complete yet simple metanorm, hardly exhausts what is possible. And in many circumstances it is sensible to allow for nuances that CNF ignores. For example, higher-order violations may be less important than lower-order ones: A's cheating B in a trade may matter more than C's continuing to cooperate with the renegade A. Hence it might be collectively desirable for the metanorm to recognize such distinctions and prescribe punishments accordingly: the more serious the violation, the heavier the sanction. (Of course, all punishments must be effective in the sense of satisfying theorem 2's criterion.)

Taken together, then, theorems 3–5 show that uniformly stable strategies must encode social norms in several fundamental ways: they hold for everyone in the community, they always hold, and any nth-order violation of the code must be punished.

Note, however, that we have *not* claimed that all uniformly stable norms are functional in the sense of being Pareto efficient. There is a good reason for not making this claim: it is false. Efficiency is *not* an essential property of uniformly stable strategies, as we will demonstrate later.

Even CNF is not strongly stable—but this is not so bad.—So far we have only been able to show that uniformly stable strategies like CNF are weakly stable. None of our results claimed that strongly stable strategies exist. The reason is simple—such strategies do *not* exist in repeated interactions. And it is easy to see why. If all members of a group play CNF and some of them mutate into always cooperate (ALL-C; cooperate unconditionally in all periods), then both strategies will have the same

[36] Using the same setting of a population of pairwise IPDs, Bendor and Mookherjee (1990) had found that under a Nash equilibrium analysis of norms, only one-level codes are necessary. It is interesting to note that changing to an evolutionary framework and requiring uniform stability implies the necessity of metanorms. (See Axelrod [1986] for a discussion of the importance of metanorms from an evolutionary perspective.)

payoff. Indeed, absent other strategies the two will be behaviorally indistinguishable. ALL-C will be a neutral mutant of CNF, and no evolutionary process can restore the preinvasion state in which everyone plays CNF. Consequently, although CNF can ensure that it will never decrease under any evolutionary process, in some invasions the mutants cannot be eliminated. Strong stability is unattainable.

However, being only weakly stable is a much less serious problem for social strategies than it is for dyadic ones. A dyadic strategy may never recover from the random drift of a neutral mutant. For example (Young and Foster 1991), suppose a small fraction of a TFT ecology mutates into the simplest nice strategy, ALL-C. Since ALL-C is a neutral mutant, it can drift randomly in the group. Now suppose the ecology is invaded by the simplest nonnice strategy, ALL-D (defect unconditionally in all periods). Even though ALL-D's entrance makes manifest the once-latent differences between TFT and ALL-C, TFT—which does not encode a *social* norm—never punishes ALL-C for its deviation of tolerating exploitation. Thus Young and Foster's simulation (1991) of an ecology composed of TFT, ALL-C, and ALL-D showed that in the long run, random drift eventually destroyed cooperation, even if initially almost everyone played TFT. The breakdown occurred because ALL-C, at first behaviorally indistinguishable from TFT, could by chance win many converts, setting the stage for the predatory ALL-D.

In contrast, as soon as CNF detects a difference between itself and a strategy that had once been a neutral mutant, it punishes the deviant. So an ecology dominated by CNF exhibits only two patterns: either everyone obeys the code of universal cooperation, whence all strategic differences remain latent and unsanctioned, or someone deviates, revealing heretofore latent differences, and CNF attacks all overt deviators. As has long been argued (Durkheim 1933; Davis 1937; Merton 1956), there is a social function to deviance: it clarifies the rules. In our model, deviance makes manifest what had been latent. Thus, so long as CNF remains sufficiently numerous, the *social state* that it enforces is, in a sense, strongly stable.

Hence, if one were to rerun the Young and Foster simulation, replacing TFT by CNF, and if the process started out with sufficiently many players using CNF, then our analytical results imply that cooperation in this modified ecology will not break down. Cooperation will be sustained because CNF attacks ALL-C as soon as the latter fails to punish ALL-D.[37] Thus the process in which ALL-C takes over by random drift is

[37] None of the three strategies can become completely extinct in the Young-Foster simulation: there is always a small background mutation rate for each strategy. This matters substantively: the constant though rare presence of ALL-D reveals to conformity that ALL-C is nonprovocable and hence a deviant. Conceivably, CNF could

blocked. This prevents the subset of cooperative (socially nice) strategies from being weakened by the infiltration of the toothless ALL-C.

Social norms can still experience random drift, of course. In an ecology composed of CNF, TFT, and ALL-D, TFT is indistinguishable from CNF: both strategies always cooperate with themselves and with each other, and punish ALL-D from period 2 on.

FUNCTIONALISM REVISITED: MUST NORMS BE OPTIMAL?

So far all of our examples of uniformly stable norms have sustained a collectively optimal outcome. (Our running example, e.g., was that of ongoing cooperation in the IPD.) Maintaining optimality in the stable state is consistent with the strong functionalist thesis. This, however, is only a part of the analytical solution. Now it is important to find out if *sub*optimal outcomes can also be supported in equilibrium. The answer to this second part of the problem turns out to be yes. In fact, a whole raft of strategies and a wide spectrum of suboptimal states can be sustained in equilibrium. To see this, we focus once again on games of enforceable cooperation.

For any particular game, label the actions so that $v(a_1,a_1) \leq \ldots \leq v(a_m,a_m)$. Thus the most "cooperative" (pure) norm would prescribe a_m as the socially desirable action, since it yields the highest payoff for a pure strategy when it plays itself. (Earlier we referred to this as a cooperative action, denoted as a_c.) A completely noncooperative norm prescribes playing a_1. Accordingly, we will often call playing a_m as "cooperating" and a_1 as "defecting." Take the smallest r (of $r = 2, \ldots, m$) such that $v(a_r,a_r) > v(a_1,a_1)$.[38] This action, a_r, is the minimally cooperative one. For this class of games, the degree of cooperation that a strategy induces in the stable populational state is a meaningful notion. For example, in the IPD we say, following Bendor and Swistak (1997), that a stable state *supports x degrees of cooperation* if x is the limiting proportion of mutually cooperative moves as the number of periods goes to infinity. (This index gives an expected frequency of cooperative moves in a stable state.) More generally, we say that a stable state supports x degrees of a_t-cooperation ($t \geq r$) if x is the limiting proportion of periods in which players choose (a_t,a_t) as the number of periods goes to infinity. We then obtain the following result.

---

be destabilized if ALL-C invaded and then took over the ecology by random drift *before* any ALL-D invaded. Since random drift takes a long time to work, this scenario is unlikely.

[38] There must be such an r because a game of enforceable cooperation has a punishment action a_d such that $v(a_m,a_m) > \max_t v(a_t,a_d)$, which implies that $v(a_m,a_m) > v(a_1,a_1)$.

THEOREM 6.—*In a repeated game of enforceable cooperation in which the future is sufficiently important, a uniformly stable state with social strategies can support any strictly positive degree of* a_t*-cooperation.*

Thus the only unstable state is the one of complete "defection."[39] This result shows that the strong functionalist thesis does not hold in our model: *evolutionary forces do not ensure* (*Pareto*) *optimal outcomes*.

FUNCTIONALISM MODIFIED: WHY FUNCTIONAL (EFFICIENT) NORMS MIGHT PREVAIL IN THE LONG RUN

Yet it seems intuitively reasonable to expect some *quantitative* difference between the stability of a cooperative strategy such as CNF and of a strategy that usually defects with its clones. Axelrod (1984) has long argued that cooperative strategies like TFT have evolutionary advantages over noncooperative ones. Indeed, elsewhere we (Bendor and Swistak 1997) have proved this to be the case: more efficient (cooperative) strategies require smaller frequencies in order to stabilize in a group. If a similar result can be established under the assumptions of our model then cooperative strategies like CNF would be more robust in two related ways. First, a strategy that is maximally cooperative with itself—always plays a_c with its clones—would be able to invade a native that is not socially nice with a lower frequency than one required by a mutant that sometimes defects with itself. Second, once established, a socially nice native would be able to resist larger invasions. Let us therefore focus on a strategy's *stabilizing frequency*—the lowest frequency that ensures that a strategy is stable (under a specified dynamic). First, we establish the minimal stabilizing frequency of any strategy in an important subset of games of enforceable cooperation. Games in this subset have a well-defined punishment action, a_d, because either (1) there is only one action that satisfies the "swords and covenants" condition,[40] as in binary choice games such as chicken or the PD; or (2) there are several such actions, but among them there is one, call it a_p, that is clearly the best punishment since, while ensuring that the cost of imposing the punishment is as low as

[39] To see why the pure Hobbesian state is not stable, consider the IPD. Let i, the native, be *nasty*—never the first to cooperate—so that it alway defects with its clones and its neutral mutants. The invaders are STFT and some nonnasty strategy j. We know that in nontrivial games one can always design a j such that $V(\text{STFT},j) > V(i,j)$. So to be uniformly stable, i must punish STFT for doing better with j than it does. But since i is already defecting in every period against STFT, it has exhausted all possible punishments. So no version of i can work. Hence any nasty native can be beaten, and so the corresponding population state of complete defection is unstable.

[40] That is, there is only one punishment action a_d such that $\max_t v(a_t,a_d) < v(a_m,a_m)$.

	x	y	z
x	2, 2	1, -6	0, 3
y	-6, 1	0, 0	0, 1
z	3, 0	1, 0	0, 0

FIG. 3.—A game with two punishment actions: both y and z are punishments, but only z is an obvious punishment action.

possible, it gives deviants the lowest payoff;[41] that is, it maximizes the native's minimum payoff.[42] We will refer to these situations as *games with obvious punishment*. (For an example of a game with more than one punishment action but only one obvious punishment see fig. 3.) All subsequent theorems will pertain to this subset of games of enforceable cooperation. Clearly, any binary choice game that is a game of enforceable cooperation must involve obvious punishment.

To keep notation simple, we denote the following payoffs to correspond to the notation in the PD: T is the "temptation" payoff, the game's largest payoff; R is the payoff to the maximally cooperative action a_m—that is, $v(a_m,a_m)$; P_1 is the payoff of the deviant's best response to punishment (where a_p is the well-defined punishment action), while P_2 denotes the lowest payoff that a player can get when implementing punishment. (In the symmetric PD, $P_1 = P_2$, but this is not generally the case.) We use a_p to denote the best punishment action in a game with obvious punishment.

Consider now a class of strategies that generalizes CNF in the following natural way. Any strategy in this class uses the standard friend-foe categorization and embodies the following two fundamental and intuitive principles of friend-foe designations:

[41] That is, $\max_t v(a_t,a_p) \leq \max_t v(a_t,a_s)$ for all a_s.

[42] That is, $\min_t v(a_p,a_t) \geq \min_t v(a_s,a_t)$ for all a_s.

1. The foe of a friend is a foe: specifically, anyone who does not "cooperate" (play a_m) with a friend is a foe.
2. The friend of a foe is a foe. There are two manifestations of being a friend of a foe: (*a*) anyone who does not punish (play a_p toward) a foe is a foe; (*b*) anyone toward whom a foe is "too friendly" is a foe.[43]

Strategies in this class thus play a_m with friends and a_p toward foes. Anyone who violates any of the fundamental principles today is a foe tomorrow; a friend in t who obeys the fundamental principles remains a friend in $t + 1$. Prior to play everyone is considered a friend. Any strategy with these properties we will call *normatively nice and retaliatory*.

THEOREM 7.—*In any repeated game of enforceable cooperation with obvious punishment and sufficiently important future, strategies which are normatively nice and retaliatory require the smallest minimal frequency in order to be uniformly stable*.[44]

Thus normatively nice and retaliatory strategies are not only uniformly stable; they are also the most robust ones. Of all uniformly stable strategies, normatively nice and retaliatory ones require the smallest frequency to be stable under all evolutionary processes. This double robustness suggests that there is something powerful indeed about normatively nice and retaliatory strategies.

It is interesting to note that any normatively nice and retaliatory native strategy creates a world with a very simple sociometric structure: from the perspective of the native strategy, people are divided into those who are in good standing and those who are not. Everyone in good standing must be on good terms with each other, and on bad terms with those in bad standing. What emerges, then, is the distinction, critical for group formation, between "us" and "them." Hence the CNF mechanism, as implemented by normatively nice and retaliatory strategies, becomes the cornerstone of the group formation process. The resulting sociometric structure would be familiar to Heider (1958) and Lévi-Strauss (1974), among others.[45]

Finally, let us directly compare the minimal stabilizing frequencies of

[43] Specifically, if i is the focal strategy that is categorizing its partners as friends or foes, then in period t strategy i regards strategy k as having been treated in "too friendly" a manner by a foe, k^*, if $V^t(i,k^*) < V^t(k,k^*)$, where $V^t(i,k)$ denotes i's payoff, through period t, when it plays k.

[44] It might be useful to illustrate this result via a numerical example from the iterated PD. Using the payoffs from Axelrod's book ($T = 5$, $R = 3$, $P = 1$, and $S = 0$), the minimal stabilizing frequency (as δ goes to one) of a normatively nice and retaliatory strategy is two-thirds. (See the proof of theorem 7 in the appendix for a formula for the minimal stabilizing frequency.) This means that invasions of nearly one-third can be repelled. Such natives have a sizable basin of attraction.

[45] We would like to thank John Padgett for this latter point.

strategies with varying degrees of efficiency. To keep the analysis straightforward and the comparisons "clean," we will compare the efficiencies of strategies that belong to the class of CNF-like strategies: all use the recursive friend-foe construction and are unforgiving (once a foe, always a foe).

Given that the actions are labeled so that $v(a_1,a_1) \leq \cdots \leq v(a_m,a_m)$, it follows that $V(i,i)(1-\delta)$ is between $v(a_1,a_1)$ and $v(a_m,a_m)$ for any pure strategy i. We can thus write $V(i,i)(1-\delta)$ as $x \cdot v(a_m,a_m) + (1-x)v(a_1,a_1)$, where $0 \leq x \leq 1$, and so interpret x as an index of strategy i's efficiency. For a completely efficient strategy, $x = 1$; for a completely inefficient strategy, $x = 0$. For example, in the IPD, if a CNF-like native prescribes cooperating in even periods and defecting in odd ones, then $x = .5$.

To ensure meaningful comparisons, in the next result we will only compare strategies that are *maximally robust* for any given degree of efficiency. That is, we will compare strategies that are the best representatives of their class of efficiency.[46]

THEOREM 8.—*In any repeated game of enforceable cooperation with obvious punishment and sufficiently important future, the more efficient the* (*maximally robust*) *CNF-type native, the larger the maximal invasion this strategy can repel.*

For example, consider a work-shirk game, where action is effort, which can be any integer in (0, . . . , 100). Assume that complete shirking (zero effort) is a strictly dominant strategy in the stage game, while the more effort one exerts the better off is one's partner. Suppose the unique symmetric optimal action is 75, so that if both people work flat out then the result is an inefficient rat race. Provided that $v(100,100) > v(0,0)$, working flat out can be supported by a uniformly stable norm-encoding strategy of the CNF type: everyone puts in 100% effort with friends and shirks completely with foes. However, the maximal invasion that this strategy can repel is smaller than that repelled by the CNF norm, which prescribes the best symmetric effort of (75, 75).

[46] In the following result, we compare CNF-type natives. Two equally efficient CNF-type natives might not be equally robust because many games of enforceable cooperation have more than one type of feasible punishment action in the stage game, and different punishments can be differentially effective in repelling invaders. By restricting our comparison to maximally robust strategies, we are implicitly requiring that strategies use their most effective punishment action, i.e., the one that permits the repulsion of the largest possible invasion, given the native's degree of efficiency. (If there are several equally effective punishments, it does not matter which one is used.)

EXTENSIONS

Important note on the robustness of norm-governed behaviors.—Theorem 7 confers a very significant methodological benefit. In its proof we have established not only the stability of certain norms but also their "degree of robustness," that is, the precise value of the minimal frequency they require in order to stabilize in any group under any evolutionary dynamic. This means that if this threshold is not breached (invasions are smaller than the biggest that the native can repel), the native strategy will remain stable—no matter the source or nature of the mutation. Since mutations can arise from any variation in the model's parameters, including, for instance, noise and uncertainty (see below), theorem 7 establishes the stability of normatively nice and retaliatory strategies under *any small variation on the model's parameters*, one that does not produce "too many" mutants. This is a very powerful kind of robustness indeed.

Network properties.—We will now elaborate on some of the most obvious and also perhaps most interesting departures from the assumptions of our model. Up to this point we have examined norms in dense social networks. These networks are dense both in terms of interaction—everyone plays everyone else in every period—and information—everyone knows, at the end of every period, what happened in all interactions in the community. A natural question, therefore, is, What happens in networks that do not meet these conditions?

First, it can be easily seen that social norms are just as necessary in sparse networks as they are in dense ones. If we consider, for instance, an environment in which both interaction and information are sparse—each person plays only one other (randomly matched) person in each period, and the outcome of each interaction becomes known to only one third party (rather than becoming known to the entire community)—then the following result is virtually a corollary of theorem 1.

PROPOSITION 2.—*In all repeated symmetric games where the future is sufficiently important, with random matching of single partners and singleton bystanders, any pure strategy that is uniformly stable must be social.*[47]

[47] The proof of this fact essentially follows that of theorem 1, with a few minor modifications. First, note that since dyadic strategies never use information about what happened in other interactions, it is as if they function in a world with no bystander observations at all. Hence the proof of theorem 1 does not depend on third-party monitoring in any way, and so this part of proposition 2 follows immediately from theorem 1. Further, the essential logic of the theorem's proof, which turns on the existence of neutral mutants and of another mutant to which the neutral mutant is preadapted, owes nothing to the density of interaction or information. Indeed, we can make the networks arbitrarily sparse, and the result still holds.

Thus social norms are indeed necessary in a much broader set of environments than the world of theorem 1.

A separate issue concerns the existence of uniformly stable norms in sparse networks. Would, for instance, the normatively infused CNF continue to be *sufficient*? Let's consider again the effects of sparse interactions versus the effects of sparse third-party monitoring. It turns out that *sparse interaction by itself creates no qualitatively new problems for social norms*. The following example shows why this is so. Suppose that in period 1 A and B happen to pair up and A cheats B. In period 2 A and E happen to meet and they cooperate. Since this is an informationally dense community, everyone knows that E failed to punish A in period 2. Consequently, from period 3 on, any player using the strategy of CNF will punish E whenever they meet her. Thus despite the sparser interaction, it is still true that no one can get away with free riding on the cooperative code. Hence CNF remains uniformly stable (provided, of course, that the future is sufficiently important). Sparse interaction, by itself, has no qualitative impact.

Similarly, delays in transmitting information about outcomes of third-party interactions do not alter any of our results.[48] The reason is, once again, quite simple. Suppose that actor E's violation in the above example becomes known to other members of the group only after some time has passed. If the discount parameter δ is sufficiently close to 1 (i.e., if future payoffs are sufficiently important), the "payoff advantage" that E can gain while her transgression remains unknown will be offset by the loss from the ensuing punishment (assuming, of course, that it is severe [long] enough).

Sparse third-party information, however, can be damaging—naturally so, since this information is the lifeblood of social norms. To see this, note that at one extreme, when there is no third-party information at all, social norms are simply infeasible: player E cannot punish A for A's cheating B if E never learns what A did. A continuity argument suggests that when it exists but is extremely sparse, such norms should also be impossible to sustain. Either too few people will know about deviations to support the native's code, or the native strategy will be so hair-trigger sensitive, to offset the infrequency of third-party information, that the native will eventually turn on and destroy itself. Near the other end of the informational continuum, if almost everyone gets third party information then uniformly stable norms do exist. People who are "out of the loop" are, in effect, playing mutant strategies, and we already know that if there are sufficiently few deviants, CNF and other norm-encoding strat-

[48] Raub and Weesie (1990) and Buskens (1999) study the effects of informational lags about third-party interactions on the efficiency of outcomes.

egies are uniformly stable. Hopefully the breakdown of social norms is analytically well behaved in that we can parameterize the amount of third-party information so that norms are uniformly stable if and only if the parameter exceeds a threshold. In general, however, this remains a matter for further research.

Uncertainty.—In the preceding analyses, norms are enforced flawlessly because players always know how their partners had behaved in previous periods. In the real world, uncertainty intrudes in various ways. Two of the most important kinds of uncertainty concern perceptions and slippages between intentions and actions (Axelrod and Dion 1988). First, players may misperceive each other's actions. Thus A thinks B cheated C when in fact B cooperated. Incorrectly believing that B violated a code of cooperation, A punishes B. The opposite error can also occur: A, mistakenly believing that B helped C, might fail to punish B's actual transgression. The second type of uncertainty involves implementation: A intends to cooperate but—possibly due to external shocks (A's computer crashes and the report cannot be completed)—fails to do so. B, observing only what A has done and not A's intention, responds accordingly.

At the beginning of this section we have already noted that if the "noise" clouding any parameters of the game does not create "too many" mutants, then theorem 7 ensures that the native norm remains stable. This need not be true if the amount of uncertainty increases. Clearly, sufficient amounts of either type of noise, misperceptions or implementation glitches, can make the enforcement of norms—or indeed the use of any strategy—problematic (see, e.g., Downs, Rocke, and Siverson 1985; Molander 1985; Donninger 1986; Bendor 1987; Mueller 1987; Nowak and Sigmund 1989, 1990, 1992, 1993; Nowak 1990; Bendor, Kramer, and Stout 1991; Kollock 1993; and Lomborg 1996).

If, for instance, the noise is such that within a generation everyone is bound to implement an action that he or she did not intend, social norms turn out not to be necessary to stabilize behaviors—in this case dyadic strategies can be uniformly stable as well.[49] In a paradoxical twist the introduction of what is usually taken to be a problem—uncertainty—has a stabilizing effect on some dyadic strategies.

But since some of these now-stable strategies are dyadic, it is no longer

[49] A constant probability of a tremble in every period of the game implies that in an infinite game trembles will reveal all latent differences among otherwise observationally indistinguishable strategies. Thus if trembles can occur in every period, neutral mutants are eliminated. This in turn allows dyadic strategies such as ALL-D to be uniformly stable. Indeed, since neutral mutants do not exist when trembles occur, a higher performance standard can be achieved: there are dyadic strategies (such as ALL-D) that are *unique* best responses to themselves (Selten 1983) and so are strongly stable under all evolutionary processes, once sufficiently numerous (Boyd 1989).

true that under this type of uncertainty stability *requires* social strategies. Would, then, social strategies retain some other evolutionary advantage over dyadic ones? Based on some simulations that we have done (not reported here) we conjecture that they would. It seems that turning a dyadic strategy into a social one can yield two types of advantages: first, it can turn an evolutionarily unstable strategy into a stable one; second, it can increase a strategy's robustness (i.e., decrease its minimal stabilizing frequency).

Consider, for instance, the Boyd-Lorberbaum ecology, discussed at length earlier, in which a native TFT is invaded by TF2T and STFT. Recall that in the absence of noise TFT is not uniformly stable in this ecology (TF2T garners the highest payoff.) In fact, the same holds true if there is noise but trembles are sufficiently infrequent. Thus for small levels of noise dyadic TFT remains evolutionarily unstable.[50] This would no longer be true, however, if one turns the dyadic TFT into the social one. The social version of TFT uses the standard friend-foe categorization in a TFT-like way: someone is a friend tomorrow if and only if they cooperate with friends and defect against foes today. This social TFT *is* stable. In fact our simulations show that it can repel invasions in the Boyd-Lorberbaum ecology even under sizable levels of noise.[51]

The relation between norms and various kinds of uncertainty is a complex and important topic that clearly merits further research. Some issues are of obvious importance for theories of institutions and public policy. Consider, for instance, the following problem: How does increasing uncertainty affect the *relative* effectiveness of enforcing codes by centralized institutions (e.g., a formal system of police and courts) versus enforcement by the more decentralized institutions of norms? In general, we suspect that enforcement by social strategies—particularly stringent ones such as CNF—will become increasingly "brittle" as relations become noisier, communities larger, and monitoring more difficult.[52] Hence, we believe that a classical thesis about enforcing codes of behavior in complex societies—if

[50] This in turn remains true even if one modifies TFT (Sugden 1986) to enable it to cope more effectively with noise. Sugden's variant (contrite tit for tat) is well-designed to handle trembles, but it is still a dyadic strategy and so cannot enforce a code of conduct on third parties.

[51] Even when the probability of making an implementation error by each player in every period of the IPD was set at 10%, the social TFT was still able to repel all invasions in the Boyd-Lorberbaum ecology.

[52] This interaction between size and monitoring problems has been studied via Nash equilibrium analysis in, e.g., Bendor and Mookherjee (1987) and Radner (1986). The general conclusion is that, given informational imperfections, the larger the community the harder it is to sustain cooperative outcomes by decentralized regimes (e.g., trigger strategies), and centralized institutions become relatively superior.

rules in such societies are to be effectively enforced then formal institutions are required—will be supported by a careful evolutionary analysis.

APPLICATIONS AND INTERPRETATIONS

This article has tried to answer the question, Why are there norms? The core of our answer is that in demanding selection environments, social norms enjoy the crucial evolutionary advantage of replicating more easily than do other behaviors. We have also established that this evolutionary advantage of social strategies and their corresponding norms continues to hold if perturbations of the basic model do not produce "too many" deviant strategies. Hence our result is robust under small perturbations of the parameters of the model. It is not, however, universal: it does not hold for all values of the parameters. Norms do not always replicate more easily. As with any formal result, we can only expect the deductive conclusions to be observed in reality if this reality conforms to the assumptions of our model. Hence in conclusion, we would like to review the meaning of several major assumptions and provide interpretations of the model's crucial parameters.

At the most general level we should note that the cognitive requirements of using norms are fundamental to a general (transspecific) theory of strategic behavior but irrelevant for sociology as long as it is confined to humans. Clearly, social strategies require more cognitive sophistication than do dyadic ones. Studies of animal behavior have shown that, for example, primates easily meet these higher requirements.[53] But some less intelligent animals, such as bats (Wilkinson 1984), reciprocate dyadically without using third-party sanctions. Thus the evolution of social norms depends on the evolution of big brains. Because all normal humans satisfy this condition, this parameter is a constant in the study of human societies and hence irrelevant in this domain. However, it *would* matter when studying the different social systems of mammals, for example.

A second general observation is that our models assume that the primary form of interaction involves two people.[54] This is a substantive assumption which may not be adequate in many situations. Such is the case with common dilemmas. The problem of collective goods has a different structure and one that is more appropriately represented as an N-person game. Such games have a different formal and conceptual structure

[53] Chimpanzees, e.g., form intricate coalitions that involve third-party interventions (de Waal 1982).

[54] Our deductive conclusion is that players will be linked to others by social strategies that impose third-party sanctions, but the basic building blocks in our model, as in Homans (1950), are two-person interactions.

in which the notion of third-party sanctions, as it was used in this article, would have to be redefined and reanalyzed.

To gain insight into the more specific assumptions embedded in our theory, consider the following hypothetical application. Assume, for instance, that we want to understand why community A has converged on prosperous cooperation and thrives economically while community B, despite many underlying similarities, is burdened with defections and ensuing economic stagnation. First, we should note that theorems 2–8 all apply to games of enforceable cooperation. Hence if we want to use these results to draw conclusions about behaviors in communities A and B, these conclusions must be limited to behaviors that can be modeled as games of enforceable cooperation. Second, since our evolutionary model allows only for an ϵ amount of noise and requires sufficiently large values of δ, it is best suited for modeling small communities where gossip is an efficient source of information and probabilities of future interactions are high.

But restrictions on the types of communities and interactions among players are not the only necessary caveats. Other conditions must hold as well. For instance, since all our theorems specify equilibrium conditions they should be applied only to communities that have attained stable states, that is, remained stable for a reasonably long time. Similarly our analysis proclaims conditions that are necessary and/or sufficient to obtain uniform stability—that is, stability under *all* types of evolutionary dynamics. Yet A and B can be stable in a weaker sense: they could be stable under *some* but not *all* evolutionary processes. (For example, elsewhere we have shown [Bendor and Swistak 1997] that if conformity incentives are sufficiently strong, then dyadic strategies such as TFT can be stable under certain evolutionary dynamics, even though certain other dynamics would destabilize TFT—and any other—dyadic strategy.)

If we can successfully assume that all of the above conditions hold, our theorems can finally be used to make a number of interesting predictions. First, theorem 1 and its corollary tell us that both the efficient equilibrium of A and the inefficient equilibrium of B must be supported by social norms. Neither a cooperative state nor a defective one can, in the long term, be supported by dyadic rewards and punishments—third-party sanctions are necessary to keep them stable. These social norms, as theorems 3, 4, and 5 tell us, must be perpetually social (they can never become dyadic), comprehensively social (no players can be excluded from the group enforcement), and thorough (they have to punish deviations of all orders).[55]

[55] As noted earlier, however, it is *not* required that these social norms be unforgiving; this feature of CNF is not essential.

The fact that uniform stability is not possible without a social structure of interactions is perhaps the most important conclusion. (It is also interesting that social structure is needed regardless of the efficiency of the group's output.)

Second, theorem 6 tells us that both the efficient state of community A and the inefficient state of community B can be explained as uniformly stable equilibria. In fact our theory predicts that any state, from the state of almost pure defection to the state of pure cooperation, can be realized as a uniformly stable equilibrium.

Third, uniform stability means that both communities are robustly stable. For instance, if we had the power to change players' incentives and push them all in the direction of *Homo economicus*, this change would not affect the equilibrium. Thus, for example, if the defective norms that imprisoned Banfield's impoverished southern Italian town (1958) had been robustly stable, then even if the citizens' values had drifted in the direction of *Homo economicus* the town's inefficient equilibrium would have persisted. Changing values would not have sufficed to destabilize that steady state of poverty. Hence, by telling us what kinds of changes will *not* work, such considerations can help us design policy mechanisms that can overturn inefficient, defection-ridden equilibria (Stinchcombe 1997).

Fourth, despite their qualitative similarity—both A and B being robustly stable—these communities exhibit an important *quantitative* dissimilarity. This difference is revealed by theorems 7 and 8, which tell us that the more cooperative a strategy (other properties equal) the lower the frequency it needs to be stable, and the lowest stabilizing frequency is attained by strategies that are cooperative and retaliatory. These in turn imply that the efficient equilibrium of A is more robust than the less efficient equilibrium of B: A will be able to resist larger invasions of mutant behaviors than would B. In other words, to maintain an inefficient equilibrium a community has to be better isolated from external shocks (immigration, cultural diffusion, etc.) than a community in an efficient equilibrium. Note that this conclusion has policy implications for the possibilities of overturning certain types of collectively undesirable equilibria.

We hope the above discussion provides a better understanding of the concepts and assumptions that underlie the deductive results of this article. We also believe that our results throw new light on some old but always vital and controversial debates about how social and economic forces are related to each other. One can clearly see, for instance, that our hypothetical example of two communities—one enjoying a good equilibrium, the other mired in a bad one—resembles the approach taken by, for example, Putnam (1993), among others, in his argument that differences

in social capital can explain the economic success of Northern Italy and the failure of the South. Our findings provide new insights and new interpretations for such issues. Indeed, Coleman's concept of social capital receives a new interpretation from the robust, efficient equilibria attained by normatively nice and retaliatory strategies. As the phrasing of his intriguing idea—the juxtaposition of "social" with "capital"—suggests, *Homo sociologicus* and *Homo economicus* must work together if societies are to reach their potential.

APPENDIX

Proofs

Theorem 2

It is easy to see that if a stage game is one of enforceable cooperation then there exist uniformly stable strategies. Consider, for instance, the strategy of CNF, as defined in the text. Take now any group $E = \{(i,p_i),(j_1,p_{j_1}), \dots ,(j_N,p_{j_N})\}$, where i denotes CNF. For any strategy j_k ($k = 1, \dots ,N$), either j_k's payoff is identical to i's (for all values of δ and p_i) or i defects with j_k from some period on. In this second case we have $V(i,i) > V(j_k,i)$ for sufficiently large δ, which implies that $V(i) > V(j_k)$ if p_i is sufficiently large. Thus, for sufficiently large p_i there is a δ_0 such that in all groups where $\delta > \delta_0$ we have $V(i) \geq V(j_k)$ which implies that i is uniformly stable.

To show that if there is a pure uniformly stable strategy in a nontrivial repeated game with sufficiently large δ then the stage game has to be a game of enforceable cooperation is more complex. Assume, by contradiction, that there exists a game which has a pure uniformly stable strategy i and yet it is not a game of enforceable cooperation. Since the stage game is not a game of enforceable cooperation there is no effective punishment action, that is, for any distinct actions a_r and a_s we have

$$v(a_r,a_r) \leq \max_t v(a_t,a_s). \tag{A1}$$

Suppose now that strategy i is uniformly stable for sufficiently high p_i and δ and consider the following ecology $E = \{(i,p_i),(j,p_j),(k,p_k)\}$. In this ecology the mutant k is constructed so as to give the mutant j the highest possible payoff while ensuring that the native i does not obtain this payoff when it plays k. This will allow j to beat i, whence i cannot be uniformly stable. So we construct strategies j and k as follows.

First, take j to be such that $V(i,i) \leq V(j,i)$. Note that since i is a pure strategy and (A1) holds, it is possible to obtain such a j. Before we construct strategy k and refine our construction of j, let's permute the m actions of the stage game so that $v(a_1,a_1) \geq \cdots \geq v(a_m,a_m)$. Thus if any of

the diagonal payoffs equal the stage game's maximal payoff of $\max_{s,t} v(a_s,a_t)$, then $v(a_1,a_1)$ must do so. In general, however, there may be more than one action which yields the game's maximal payoff on the diagonal. Thus the payoffs on the diagonal must belong to one of two distinct subsets, maximal and nonmaximal, that is, there is an integer w, where $0 \leq w < m$, such that $v(a_r,a_r) = \max_{s,t} v(a_s,a_t)$ for all $r = 1, \ldots, w$ and $v(a_r,a_r) < \max_{s,t} v(a_s,a_t)$ for all $r = w + 1, \ldots, m$. (If no diagonal payoff is maximal, then $w = 0$.)

Note that if for any $r \leq w$ $v(a_r,a_r) = v(a_r,a_t)$ for all t then Row would get the maximal payoff no matter what Column did, that is, the stage game would be trivial. Since we have assumed that it is *not* trivial, for any $r \leq w$ there must be a t such that

$$v(a_r,a_r) > v(a_r,a_t) \tag{A2}$$

For any $r \leq w$ denote an action a_t for which (A2) holds as a_{t_r}. Now consider the following mixed strategy:

$$k^* = (1/m)a_{t_1} + \cdots + (1/m)a_{t_w} + (1/m)a_{w+1} + \cdots + (1/m)a_m,$$

which means that k^* plays $a_{t_1}, \ldots, a_{t_w}, a_{w+1}, \ldots, a_m$ each with probability $1/m$. (This construction ensures that no strategy which plays k^* can get the stage game's maximal payoff.) Let a_x and a_y, where $1 \leq x,y \leq m$ be such that $v(a_x,a_y) = \max_{r,s} v(a_r,a_s)$. Now, construct strategy k and refine strategy j as follows. In period 1 k plays any action that is different than the action of i and j in period 1. In period 2 j plays any action toward k which is different than what j plays against i. Now, from period 3 on k plays k^* against i and a_y against j, whereas j plays a_x against k. Note that in any period $t > 2$, $v^t(j,k) = \max_{r,s} v(a_r,a_s)$ and, given the construction of k, $v^t(i,k) < \max_{r,s} v(a_r,a_s)$, where $v^t(i,j)$ denotes payoff in period t in a game between i and j. Thus, for sufficiently high δ we have $V(j,k) > V(i,k)$. For any p_i and sufficiently large δ as p_j converges to zero, $V(i)$ converges to $p_iV(i,i) + (1 - p_i)V(i,k)$ whereas $V(j)$ converges (the left-hand-side limit only) to $p_iV(j,i) + (1 - p_i)V(j,k)$. Since $V(j,k) > V(i,k)$, for any p_i we can find p_j and p_k such that $V(i) < V(j)$ for sufficiently large δ's. This, however, means that i is *not* uniformly stable, and so we have a contradiction. Q.E.D.

Theorem 3

Suppose that strategy i is not perpetually social. Then there is some period t after which it becomes dyadic. Then let there be an arbitrarily small invasion by strategies j and k, where these strategies are designed as follows. Both are neutral mutants of i for periods $1, \ldots, t$. After t, strategy

j continues to behave as i does in all encounters with i and itself; hence $V(i,i) = V(i,j) = V(j,i) = V(j,j)$. However, k differentiates itself from the native in $t + 1$, and in $t + 2$ j behaves differently toward k than i does. This differentiation permits k to play with j differently than it does with i, ensuring (for any nontrivial game) that $V(j,k) > V(i,k)$, and hence $V(j) > V(i)$, for sufficiently large δ. Q.E.D.

Theorem 4

Suppose that strategy i is not comprehensively social. Then there is a set of strategies $\{i,j,k\ldots\}$ and period t in the game such that in all periods following t i's actions toward all strategies in the group are independent of their interactions with j. Assume now that strategy k is a neutral mutant of i up to and including period t. From period $t + 1$ on k behaves the same way i does toward all strategies in the group, except for strategy j. More specifically, consider k and j as in the proof of theorem 3. Strategy i will score the same as j with all strategies in the group except for k. And since j scores higher with k than i does, for sufficiently large δ we will get $V(j) > V(i)$. But this means that i is not uniformly stable. Q.E.D.

Theorem 5

Suppose in any game of enforceable cooperation, a strategy i does not have a metanorm structure. Then there must be some nth-order deviation that i does not punish. Construct a mutant i' that is identical to i except that i' commits nth-order deviations. Now invade i by a combination of i' plus a set of $n - 1$ distinct mutant strategies, where j_1 commits a first-order deviation, mutant j_2 commits a second-order deviation (does not punish first-order deviations), and so on until mutant j_{n-1}, which does not punish $n -$ second-order deviations. All of these strategies respond to punishment by playing a_d. The native, i, punishes all strategies (i.e., $j_1, \ldots, j_{n-1}$), while i' punishes all except j_{n-1}, with whom it continues in a "cooperative" relationship where both play a_c in every period. Thus with strategy j_{n-1}, the native gets only $v(a_d,a_d)$ as a per period average payoff, while the mutant i' gets $v(a_c,a_c)$, which must be higher. Since by construction i' behaves identically to i with all other strategies, it gets the same payoff with them as i does, whence $V(i') > V(i)$ for sufficiently large δ. Q.E.D.

Theorem 6

We will say that a pure strategy i is *ecology neutral* with i^* in ecology E, where $i,i^* \in E$, if an action taken by i in any period against any opponent

$j \in E$ is the same as i^*'s action against j. Two strategies which are ecology neutral in E are behaviorally indistinguishable in E since they act the same against all other strategies in E. Consider now an infinite sequence of zeros and ones. Take the initial n elements of the sequence and denote by $n(1)$ the number of ones among them. Any real number $0 \leq r \leq 1$ can be represented (usually in more than one way) by the following limit: $\lim_{n\to\infty}[n(1)/n]$. Take now a pure strategy i, which "cooperates" (plays a_t) only in periods corresponding to the element "1" in the infinite series as long as its opponent j does the same in a game with i as well as in all other pairwise games in the ecology (i.e., if j is ecology neutral with i); if j moves differently in any period of the game, i responds by "defecting" (playing some punishment action) in all periods that follow. By the definition of i, i supports r degrees of a_t-cooperation when universal in the population.

If $r > 0$, then a strategy i that supports r degrees of a_t-cooperation must cooperate when playing its clone in an infinite number of periods. If a strategy j is not ecology-neutral with i, then j moves differently than i against some strategy k in some period t of the game. By the definition of i, from period $t + 1$ on, i will support r degrees of a_t-cooperation when playing another i while "defecting" with j in all moves. Thus for sufficiently high δ we will have $V(i,i) > V(j,i)$, which ensures that i is uniformly stable. Q.E.D.

Theorem 7

We will first show that, in any repeated game of enforceable cooperation with obvious punishment and sufficiently high δ, the minimal frequency that stabilizes a strategy under any evolutionary dynamic is $(T - P_2)/(T + R - P_1 - P_2)$. Assume, by contradiction, that there is a strategy i with the minimal stabilizing frequency $p_0 = [(T - P_2)/(T + R - P_1 - P_2)] - \epsilon$, where $\epsilon > 0$. Consider $E = \{(i,p_0),(j,p_1),(k,p_2)\}$, where $p_0 = [(T - P_2)/(T + R - P_1 - P_2)] - \epsilon$. Let j and k be such that after some initial trigger moves, j always defects with i and k and always cooperates with other j's, while k always defects with i and cooperates with j. For example, if i cooperates in the first period take j and k as follows: j defects in periods 1 and 2 (unconditionally), and then cooperates with all strategies which defected in the first two periods and defects unconditionally with all other strategies; k defects in period 1 and then cooperates unconditionally with all strategies that defected in period 1 and defects unconditionally otherwise. (A similar construction is possible when i defects in period 1.)

For a moment we do not assume anything about the frequencies of j and k in the population, other than, obviously, $p_0 + p_1 + p_2 = 1$. In such an ecology, $\min \lim_{\delta\to 1} V(i)(1 - \delta) = p_0R + p_1P_2 + p_2P_2$ and $\max \lim_{\delta\to 1}$

$V(j)(1-\delta) = p_0P_1 + p_1R + p_2T$. For i to be stable under all evolutionary processes it is necessary that $\max \lim_{\delta\to 1} V(i)(1-\delta) \geq \min \lim_{\delta\to 1} V(j)(1-\delta)$, that is, $p_0R + p_1P_2 + p_2P_2 \geq p_0P_1 + p_1R + p_2T$. Denoting $p_1 = \varepsilon$, we can write this inequality as

$$p_0 > [(T-P_2)/(T+R-P_1-P_2)] - [\varepsilon(T-R)/(T+R-P_1-P_2)].$$

We have assumed, however, that $p_0 = [(T-P_2)/(T+R-P_1-P_2)] - \epsilon$, which gives us $\epsilon \leq \varepsilon[(T-R)/(T+R-P_1-P_2)]$. Since nothing was assumed about the value of ε, we can clearly take ε small enough to get $\epsilon > \varepsilon[(T-R)/(T+R-P_1-P_2)]$ which contradicts i's minimal stabilizing frequency being $[(T-P_2)/(T+R-P_1-P_2)] - \epsilon$.

We will now prove that in any repeated game of enforceable cooperation with obvious punishment and sufficiently high δ, strategies which are normatively nice and retaliatory require the smallest minimal frequency of $[(T-P_2)/(T+R-P_1-P_2)]$ to be uniformly stable.

Take any group $E = \{(j_1,p_1),(j_2,p_2),\dots,(j_N,p_N)\}$ in which j_1 is normatively nice and retaliatory. Consider any j_m $(m = 2,\dots,N)$. We will examine the following two cases: (1) j_1 cooperates infinitely often with j_m, that is, for each period n there is a period k, $k > n$, such that j_1 cooperates with j_m in period k; and (2) j_1 cooperates finitely often with j_m, that is, there exists a period k such that in each period n, $n > k$, j_1 defects with j_m.

Case 1.—If case 1 holds, then take any period k such that j_1 cooperates with j_m in period $k+1$. Since j_1 is normatively retaliatory, we get $V^k(j_1,j_t) \geq V^k(j_m,j_t)$, for all j_t $(t = 2,\dots,N)$. Moreover, since for every n there is $k > n$ such that j_1 cooperates with j_m in period $k+1$, there is an infinite series $k_1,k_2,\dots$, such that for every k_r, $V^{k_r}(j_1,j_t) \geq V^{k_r}(j_m,j_t)$. Consequently, $V(j_1,j_t) \geq V(j_m,j_t)$. Thus $V(j_1) \geq V(j_m)$ for all values of δ, which completes the proof of case 1.

Case 2.—Consider, again, an arbitrary j_m $(m = 2,\dots,N)$. In case 2 there is a period $k+1$ such that in all periods n, $n > k$, j_1 defects with j_m. Let's decompose the total payoff V into $V = V^k + V^{k\rightarrow}$, where V^k is the payoff after the first k periods and $V^{k\rightarrow}$ is the continuation payoff (in the remaining infinite part of the game.) Consider the second part of the game, that is, the $V^{k\rightarrow}$'s. In this part j_1 defects with j_m in all periods. From case 1, we know that for any strategy j_t if j_1 cooperates with j_t in infinitely many periods then $\min[V(j_1,j_t) - V(j_m,j_t)] \geq 0$; if, however, j_1 defects with j_t from some period on, then $\min[V(j_1,j_t) - V(j_m,j_t)] = \min V(j_1,j_t) - \max V(j_m,j_t) = P_2 - T < 0$. Since this second expression is smaller, $\min[V(j_1,j_t) - V(j_m,j_t)] = P_2 - T$. Thus,

$$\min[V^{k\rightarrow}(j_1) - V^{k\rightarrow}(j_m)] = \min[V^{k\rightarrow}(j_1,j_1) - V^{k\rightarrow}(j_m,j_1)]p_1 + \min\left[\sum_{t=2}^{N}[V^{k\rightarrow}(j_1,j_t) - V^{k\rightarrow}(j_m,j_t)]p_t\right] = p_1(R - P_1) + (1 - p_1)(P_2 - T).$$

Considering now that $\lim_{\delta\rightarrow 1} V^{k\rightarrow}(j^*) = \lim_{\delta\rightarrow 1} V(j^*)$ for any strategy j^*, $V(j_1) - V(j_m)$ approaches 0 as δ approaches 1 and $p_1 = (T - P_2)/(T + R - P_1 - P_2)$, which is equivalent to saying that the minimal stabilizing frequency of j_1 approaches $(T - P_2)/(T + R - P_1 - P_2)$. This completes the proof of case 2 and theorem 7. Q.E.D.

Theorem 8

Since a maximally robust native must use the most effective obvious punishment action, without loss of generality we restrict attention to the case where all the strategies being compared use the obvious punishment action a_p. Suppose now that a CNF native strategy i has level of efficiency x, where $\lim_{\delta\rightarrow 1} V(i)(1 - \delta) = [xv(a_m,a_m) + (1 - x)v(a_1,a_1)]$.

Consider an ecology where CNF would be most vulnerable to invasion. Such a "test" invasion for a CNF native, i, is one in which there are two mutants, say j and k, where $\epsilon_j + \epsilon_k = \epsilon$ and the invaders are constructed so as to boost the fitness of one of the strategies (say, j) and to reduce the fitness of the native as much as possible, in order to create a winning mutant ($V(j) > V(i)$) with the smallest possible invasion size. Thus j and k are constructed as follows: j plays a best response to the native's punishment and plays a_m with itself. Mutant k, after "recognizing" j, plays so as to maximize j's pairwise score; thus $\lim_{\delta\rightarrow 1} V(j,k)(1 - \delta) = T$. Meanwhile, k responds to i's punishment so as to minimize $v(a_p,a_t)$. Given this construction, $V(j)$ is increasing as $\epsilon_j \rightarrow 0$, so we consider invasions in which the proportion of j-mutants is negligible.

For such invasions,

$$\lim_{\epsilon_j\rightarrow 0}\lim_{\delta\rightarrow 1} V(i)(1 - \delta) = (1 - \epsilon)[xv(a_m,a_m) + (1 - x)v(a_1,a_1)] + \epsilon P_2,$$

whereas $\lim_{\epsilon_j\rightarrow 0}\lim_{\delta\rightarrow 1} V(j)(1 - \delta) = (1 - \epsilon)P_1 + \epsilon T$. Since $v(a_m,a_m) > v(a_1,a_1)$, $V(i)$ is increasing in x, while $V(j)$ is unaffected. Hence as x increases it follows that ϵ, the invasion size, must increase in order to ensure that $V(j) > V(i)$. This means that the more efficient i is, the larger the maximal invasion it can repel under any evolutionary dynamic. Q.E.D.

REFERENCES

Aumann, Robert. 1987. "Game Theory." In *The New Palgrave: A Dictionary of Economics*. New York: Norton.

Axelrod, Robert. 1981. "The Emergence of Cooperation among Egoists." *American Political Science Review* 75:306–18.

———. 1984. *The Evolution of Cooperation*. New York: Basic Books.

———. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80:1095–1111.

Axelrod, Robert, and Douglas Dion. 1988. "The Further Evolution of Cooperation." *Science* 242 (December 9): 1385–90.

Axelrod, Robert, and William Hamilton. 1981. "The Evolution of Cooperation." *Science* 211 (March 27): 1390–96.

Banfield, Edward. 1958. *The Moral Basis of a Backward Society*. New York: Free Press.

Bendor, Jonathan. 1987. "In Good Times and Bad: Reciprocity in an Uncertain World." *American Journal of Political Science* 31:531–58.

Bendor, Jonathan, Roderick Kramer, and Suzanne Stout. 1991. "When in Doubt . . . Cooperation in a Noisy Prisoner's Dilemma." *Journal of Conflict Resolution* 35: 691–719.

Bendor, Jonathan, and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action." *American Political Science Review* 81:133–47.

———. 1990. "Norms, Third-Party Sanctions, and Cooperation." *Journal of Law, Economics, and Organization* 6:33–63.

Bendor, Jonathan, and Piotr Swistak. 1996. "The Controversy about the Evolution of Cooperation and the Evolutionary Roots of Social Institutions." Pp. 113–35 in *Social Agency*, edited by W. Gasparski, M. Mlicki, and B. Banathy. New Brunswick, N.J.: Transaction Publishers.

———. 1997. "The Evolutionary Stability of Cooperation." *American Political Science Review* 91:290–307.

———. 1998. "Evolutionary Equilibria: Characterization Theorems and Their Implications." *Theory and Decision* 45:99–159.

———. 1999. "The Evolution of Universalistic and Ascriptive Norms." Working paper. Stanford University.

———. 2000. "A Theory of Social Behavior and a Solution to the Riddle of Social Norms." Working paper. Stanford University.

Bernheim, B. D. 1984. "Rationalizable Strategic Behavior." *Econometrica* 52:1007–28.

Bicchieri, Cristina. 1993. *Rationality and Coordination.* New York: Cambridge University Press.

Binmore, Kenneth. 1992. *Fun and Games: A Text on Game Theory*. Lexington, Mass.: D.C.: Heath.

Binmore, Kenneth, and Larry Samuelson. 1994. "An Economist's Perspective on the Evolution of Norms." *Journal of Institutional and Theoretical Economics* 150:45–63.

Blake, Judith, and Kingsley Davis. 1964. "Norms, Values, and Sanctions." In *Handbook of Modern Sociology,* edited by Robert Faris. Chicago: Rand McNally.

Blau, Peter M. 1964. *Exchange and Power in Social Life*. New York: Wiley.

Bomze, Immanuel, and Benedikt Potscher. 1989. *Game Theoretical Foundations of Evolutionary Stability*. New York: Springer-Verlag.

Bomze, I. M., and Eric van Damme. 1992. "A Dynamical Characterization of Evolutionarily Stable States." *Annals of Operations Research* 37:229–44.

Boyd, Robert. 1989. "Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma Game." *Journal of Theoretical Biology* 136:47–56.

Boyd, Robert, and Jeffrey Lorberbaum. 1987. "No Pure Strategy Is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game." *Nature* 327 (May 7): 58–59.

Boyd, Robert, and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.
Buskens, Vincent. 1999. *Social Networks and Trust*. Amsterdam: Thela Thesis.
Cabrales, Antonio. 1993. "Stochastic Replicator Dynamics." Economics Working Paper no. 54. Barcelona: Universitat Pompeu Fabra.
Calvert, Randall. 1995. "Rational Actors, Equilibrium, and Social Institutions." In *Explaining Social Institutions*, edited by Jack Knight and Itai Sened. Ann Arbor: University of Michigan Press.
Coleman, James S. 1986. "Social Theory, Social Research, and a Theory of Action." *American Journal of Sociology* 91:1309–35.
———. 1990*a*. *Foundations of Social Theory*. Cambridge, Mass.: Harvard University Press.
———. 1990*b*. "Norm-Generating Structures." In *The Limits of Rationality*, edited by Karen Cook and Margaret Levi. Chicago: University of Chicago Press.
Coleman, James, and Thomas Fararo, eds. 1992. *Rational Choice Theory: Advocacy and Critique*. Newbury Park, Calif.: Sage Publications.
Cressman, Ross. 1992. *The Stability Concept of Evolutionary Game Theory.* Berlin: Springer-Verlag.
Davis, James. 1967. "Clustering and Structural Balance in Graphs." *Human Relations* 20:181–87.
Davis, Kingsley. 1937. "The Sociology of Prostitution." *American Sociological Review* 2:744–55.
Dawkins, Richard. 1989. *The Selfish Gene*, 2d ed. Oxford: Oxford University Press.
de Waal, Frans. 1982. *Chimpanzee Politics: Power and Sex among Apes.* New York: Harper & Row.
Donninger, Christian. 1986. "Is It Always Efficient to Be Nice? A Computer Simulation of Axelrod's Computer Tournament." Pp. 123–34 in *Paradoxical Effects of Social Behavior*, edited by A. Diekmann and P. Mitter. Heidelberg: Physica-Werlag.
Downs, George, David Rocke, and Randolph Siverson. 1985. "Arms Races and Cooperation." *World Politics* 38:118–46.
Durkheim, Émile. 1933. *The Division of Labor in Society*. New York: Macmillan.
Eggertsson, Thrainn. 2001. "Norms in Economics with Special Reference to Economic Development." In *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp. New York: Russell Sage Foundation.
Ellickson, Robert. 1991. *Order without Law.* Cambridge, Mass.: Harvard University Press.
Elster, Jon. 1983. *Explaining Technical Change.* Cambridge: Cambridge University Press.
——— .1989*a*. *The Cement of Society.* Cambridge: Cambridge University Press.
——— .1989*b*. *Nuts and Bolts for the Social Sciences.* Cambridge: Cambridge University Press.
Flache, Andreas, and Michael Macy. 1996. "The Weakness of Strong Ties: Collective Action Failure in a Highly Cohesive Group." *Journal of Mathematical Sociology* 21:3–28.
Frank, Robert. 1988. *Passions within Reason*. New York: Norton.
Friedman, Daniel. 1991. "Evolutionary Games in Economics." *Econometrica* 59: 637–66.
Fudenberg, Drew, and David K. Levine. 1998. *The Theory of Learning in Games.* Cambridge: MIT Press.
Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory*. Cambridge, Mass.: MIT Press.
Gale, John, Kenneth Binmore, and Larry Samuelson. 1995. "Learning to be Imperfect: The Ultimatum Game." *Games and Economic Behavior* 8:56–90.
Gibbs, Jack. 1968. "The Study of Norms." In *International Encyclopedia of the Social Sciences*, vol. 11. Edited by David Sills. New York: Macmillan.

Hardin, Russell. 1995. *One for All.* Princeton, N.J.: Princeton University Press.
Hechter, Michael. 1987. *Principles of Group Solidarity*. Berkeley: University of California Press.
Hechter, Michael, and Satoshi Kanazawa. 1997. "Sociological Rational Choice Theory." *Annual Review of Sociology* 23:191–214.
Heckathorn, Douglas. 1988. "Collective Sanctions and the Emergence of Prisoner's Dilemma Norms." *American Journal of Sociology* 94:535–62.
———. 1990. "Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Social Control." *American Sociological Review* 55:366–84.
Heider, Fritz. 1958. *The Psychology of Interpersonal Relations.* New York: John Wiley & Sons.
Hines, W. G. S. 1987. "Evolutionary Stable Strategies: A Review of Basic Theory." *Theoretical Population Biology* 31:195–272.
Hofbauer, Josef, and Karl Sigmund. 1988. *The Theory of Evolution and Dynamical Systems*. Cambridge: Cambridge University Press.
Homans, George C. 1950. *The Human Group*. New York: Harcourt, Brace.
———. 1961. *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace.
Horne, Christine. 2001. "Sociological Perspectives on the Emergence of Norms." In *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp. New York: Russell Sage Foundation.
Kollock, Peter. 1993. "An Eye for an Eye Leaves Everyone Blind: Cooperation and Accounting Systems." *American Sociological Review* 58:768–86.
Lévi-Strauss, Claude. 1974. *Structural Anthropology.* New York: Basic Books.
Lomborg, Bjorn. 1996. "Nucleus and Shield: The Evolution of Social Structure in the Iterated Prisoner's Dilemma." *American Sociological Review* 61:278–307.
Macy, Michael W. 1993. "Backward-Looking Social Control." *American Sociological Review* 58:819–36.
Macy, Michael W., and John Skvoretz. 1998. "The Evolution of Trust and Cooperation between Strangers: A Computational Model." *American Sociological Review* 63: 638–60.
Mailath, George J. 1992. "Introduction: Symposium on Evolutionary Game Theory." *Journal of Economic Theory* 57:259–77.
March, James G., and Herbert A. Simon. 1958. *Organizations*. New York: Wiley.
Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
Maynard Smith, John, and G. Price. 1973. "The Logic of Animal Conflict." *Nature* 246:15–18.
Merton, Robert. 1956. *Social Theory and Social Structure*. New York: Free Press.
Molander, Per. 1985. "The Optimal Level of Generosity in a Selfish, Uncertain Environment." *Journal of Conflict Resolution* 29:611–18.
Moore, Wilbert E. 1978. "Functionalism." In *A History of Sociological Analysis*, edited by Tom Bottomore and Robert Nisbet. New York: Basic Books.
Mueller, Ulrich. 1987. "Optimal Retaliation for Optimal Cooperation." *Journal of Conflict Resolution* 31:692–724.
Myerson, Roger. 1991. *Game Theory*. Cambridge, Mass.: Harvard University Press.
Nowak, Martin. 1990. "Stochastic Strategies in the Prisoner's Dilemma." *Theoretical Population Biology* 38:93–112.
Nowak, Martin, and Karl Sigmund. 1989. "Oscillations in the Evolution of Reciprocity." *Journal of Theoretical Biology* 137:21–26.
Nowak, Martin, and Karl Sigmund. 1990. "The Evolution of Stochastic Strategies in the Prisoner's Dilemma." *Acta Applicandae Mathematicae* 20:247–65.
———. 1992. "Tit for Tat in Heterogeneous Populations." *Nature* 355:250–53.
———. 1993. "A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game." *Nature* 364:56–58.

Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action." *American Journal of Sociology* 85:1356–75.
Opp, Karl-Dieter. 1982. "The Evolutionary Emergence of Norms." *British Journal of Social Psychology* 21:139–49.
Pearce, D. G. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection." *Econometrica* 52:1029–1050.
Putnam, Robert D. 1993. *Making Democracy Work.* Princeton, N.J.: Princeton University Press.
Radner, Roy. 1986. "Repeated Partnership Games with Imperfect Monitoring and No Discounting." *Review of Economic Studies* 53:43–57.
Raub, Werner, and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96: 626–54.
Rasmusen, Eric. 1989. *Games and Information*. Oxford: Basil Blackwell.
Rubinstein, Ariel. 1991. "Comments on the Interpretation of Game Theory." *Econometrica* 59:909–24.
Samuelson, Larry. 1993. "Recent Advances in Evolutionary Economics: Comments." *Economics Letters* 42:313–19.
———. 1998. *Evolutionary Games and Equlibrium Selection.* Cambridge: MIT Press.
Schotter, Andrew. 1981. *The Economic Theory of Institutions.* Cambridge: Cambridge University Press.
Selten, Reinhard. 1983. "Evolutionary Stability in Extensive 2-Person Games." *Mathematical Social Sciences* 5:269–363.
———. 1991. "Evolution, Learning, and Economic Behavior." *Games and Economic Behavior* 3:3–24.
Simon, Herbert A. 1957. *Models of Man*. New York: Wiley.
Sobel, Joel. 1993. "Evolutionary Stability and Efficiency." *Economic Letters* 42:301–12.
Stinchcombe, Arthur. 1968. *Constructing Social Theories.* New York: Harcourt, Brace & World.
———. 1997. "On the Virtues of the Old Institutionalism." *Annual Review of Sociology* 23:1–18.
Sugden, Robert. 1986. *The Economics of Rights, Co-operation and Welfare*. Oxford: Basil Blackwell.
———. 1989. "Spontaneous Order." *Journal of Economic Perspectives* 3:85–97.
Taylor, Howard. 1970. *Balance in Small Groups.* New York: Van Nostrand Reinhold.
Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46:35–57.
Ullmann-Margalit, Edna. 1977. *The Emergence of Norms.* Oxford: Oxford University Press.
van Damme, Eric. 1987. *Stability and Perfection of Nash Equilibria.* Berlin: Springer-Verlag.
Vega-Redondo, Fernando. 1996. *Evolution, Games, and Economic Behavior.* Oxford: Oxford University Press.
Vincent, Thomas L., and Joel S. Brown. 1988. "The Evolution of ESS Theory." *Annual Review of Ecology and Systematics* 19:423–43.
Voss, Thomas, and Martin Abraham. 1999. "Rational Choice Theory in Sociology: A Survey." Discussion paper. University of Leipzig, Department of Sociology.
———. 2001. "Game Theoretical Perspectives on the Emergence of Social Norms." In *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp. New York: Russell Sage Foundation.
Warneryd, Karl. 1993. "Cheap Talk, Coordination, and Evolutionary Stability." *Games and Economic Behavior* 5:532–46.
Weibull, Jorgen. 1995. *Evolutionary Game Theory*. Cambridge, Mass.: MIT Press.

Wilkinson, Gerald. 1984. "Reciprocal Food Sharing in the Vampire Bat." *Nature* 308 (March 8): 81–84.
Wrong, Dennis. 1961. "The Oversocialized Conception of Man in Modern Sociology." *American Sociological Review* 26:184–93.
Young, Peyton H. 1998. *Individual Strategy and Social Structure.* Princeton, N.J.: Princeton University Press.
Young, Peyton, and Dean Foster. 1991. "Cooperation in the Short and in the Long Run." *Games and Economic Behavior* 3:145–56.