Trust among Strangers*

Cristina Bicchieri, John Duffy, and Gil Tolle†‡

The paper presents a simulation of the dynamics of impersonal trust. It shows how a "trust and reciprocate" norm can emerge and stabilize in populations of conditional cooperators. The norm, or behavioral regularity, is not to be identified with a single strategy. It is instead supported by several conditional strategies that vary in the frequency and intensity of sanctions.

1. Introduction. Social exchanges often involve a time lag between promise and delivery. This is not just common to market exchanges, but to political exchanges as well. Markets often involve anonymous, one-shot transactions, and the working of democracies presupposes that when a citizen gives her vote she expects the representative to fulfill his part of an informal, tacit 'contract.' Both sides can benefit from an honest exchange, yet there is the potential for cheating. The motivations of those we interact with cannot be known directly, and the quality of goods and services we are offered is often unknown. If we trust, we make ourselves vulnerable to exploitation, since others' behavior is not under our control. By trust we thus mean a disposition to engage in social exchanges that involve uncertainty and vulnerability, but that are also potentially rewarding. This disposition may be grounded upon a belief in the trustworthiness of the specific agents with whom we interact, either because we directly or indirectly know about their past behavior, or else because we see that it

*Received December 2002; revised September 2003.

†To contact the authors please write to: Cristina Bicchieri, Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: cb36@andrew.cmu.edu. John Duffy, Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260; e-mail: jduffy@pitt.edu. Gilman Tolle, Computer Science, University of California, Berkeley, Berkeley, CA 94720; e-mail: get@eecs.berkeley.edu.

‡We wish to thank Jason M. Alexander and two anonymous referees for their comments and suggestions.

Philosophy of Science, 71 (July 2004) pp. 286–319. 0031-8248/2004/7101-0003$10.00

might be in their long-term interest to reward our trust, even if their short-term interests militate against rewarding it. The most interesting cases, however, are those in which it may not be in another's self-interest (however defined) to be trustworthy, we do not have personal experience with this person, group or organization, but entering into a relation (commercial, political, etc.) with it may prove extremely profitable. In situations in which we have little information and little time to gather more, trust might be better described as a disposition to engage in impersonal social exchanges, hence the name 'impersonal trust.' People may of course have expectations of trustworthiness also in these situations, but these expectations will not be grounded upon the recognition that it is in the other party's interest to be perceived as trustworthy.[1] Expectations in this case may be adaptive, meaning that past experiences will loom large in one's willingness to trust and/or reciprocate in impersonal, even anonymous exchanges. What we are modeling here, however, is not the emergence of expectations about the trustworthiness of other parties or institutions. We are rather interested in the emergence of a behavioral pattern of trusting/reciprocating.

Note that the problem of impersonal trust is a classical case of a one-sided social dilemma.[2] The standard solution to "social dilemmas" is to introduce some form of formal or informal social control. Formal controls involve the existence of 'impartial' agencies that monitor and sanction compliance with agreements we enter into. Institutional protection, however, can be costly. Monitoring and sanctioning require a complex organization, and often the very existence of such formal controls can be counterproductive (Fehr, Gachter, and Kirchsteiger 1997), in that it creates an atmosphere of distrust. Informal controls instead rest on the possibility of repeated, non-anonymous interactions. The repetition itself, with its possibilities for signaling, retaliation, and reputation formation, becomes an enforcement mechanism. Thus a network of stable exchange relationships is a source of trust, since people will prefer to transact with individuals or organizations that have a known reputation for honesty. The drawback of this solution is that transactions will be limited to a restricted network. Patron-client political exchanges, as well as the common business practice of shunning better deals in favor of established suppliers (Yamagishi and Yamagishi 1994) are examples of parochial tendencies that may ultimately backfire, if the opportunity cost of avoiding the open market becomes too great. The problem of how impersonal trust can emerge and persist is thus intertwined with the possibility of the

1. For an exhaustive discussion of trust, its meaning, and its economic and political consequences, see Dasgupta 1988, Cook 2001.

2. It is one-sided because the brunt of risk and possible loss is borne only by the trustor.

transient and anonymous social exchanges that are the foundation of market economies and democratic systems.[3] When there is an occasion for cheating, there is often the promise of great gain, and though the incidence of dishonesty may be higher among strangers than among neighbors, it is by no means a universal phenomenon.[4]

Our goal is to model how a behavioral pattern of trusting and reciprocating may develop among boundedly rational agents in a complex environment in the absence of formal or informal controls.[5] To do that, we abandon the traditional assumptions of perfect rationality, unlimited calculating capabilities, and extensive knowledge that are the backbone of classical game-theoretic models. In the complex environments we consider, computing an optimal strategy is a daunting task. What we instead model is the process through which different strategies interact and how their mix evolves. In an evolutionary model, strategies that have been relatively effective in a population become more widespread, and strategies that have been less effective become less common in the population. There is by now a vast literature on the evolution of cooperative strategies in Prisoner's dilemma games (Axelrod 1984). The kind of interaction we study is called a Trust game, and is like an alternating, one-sided Prisoner's dilemma, in that players have asymmetric roles and move sequentially, and only one of them is given the chance of cooperating or defecting at any time. Another difference between our model and those that study the emergence of cooperation in evolutionary games is that in our model strategies are not just history-contingent; they are also role-contingent, in the sense that a player must have a plan of action for each of the roles (truster or trustee) in which she may be cast.

The paper shows the results of deterministic and stochastic simulations in both one-shot and repeated versions of the Trust game. The present results show how trust and reciprocity can emerge in a population of strangers and examine whether these behaviors remain robust to a change in the size of the strategy space and/or the length of the game. An interesting result of both simulations is that many of the conditional pure strategies support trusting/reciprocating or 'cooperative' behavior, but none of them is an evolutionarily stable strategy. In fact, the concept of *evolutionarily stable*

3. Social capital refers to the norms and networks that enable collective action. A persistent problem in the social capital literature is precisely to explain how 'local' trust borne out of social networks can extend to interactions with large, anonymous groups; cf. Putnam (1993, 2000).

4. The experimental literature on "trust games" reveals a wide variety of behavior when subjects play in fixed pairs, in groups and/or with strangers. See Camerer (2003) for a survey.

5. By bounded rationality we mean that agents follow simple, not necessarily optimal rules of behavior, and this reflects limited understanding of the environment in which they operate.

strategy is not the relevant analytic concept here. The relevant analytic concept here is that of an *evolutionarily stable state,* which means a stable mix of strategies appearing in different proportions in the population.[6] What we have is a vector of strategies, or a polymorphic population, in which each player uses a pure strategy and different players may use different strategies. These populations are made of a stable majority of conditional trusting/reciprocating types; interestingly, unconditional trusting/reciprocating types can survive as a small minority only in those populations where the majority is made of conditional types.

In our model, though the underlying strategies are heterogeneous, observable actions are homogenous, i.e., trusting/reciprocating behavior is widespread. This means that an external observer would detect a behavioral regularity, and might be misled into thinking that players use the same strategy, possibly an unconditional one. A social norm is, among other things, a regular behavioral pattern. The behavior dictated by a norm, however, is usually conditional.[7] Indeed, an important difference between a social norm and an unconditional rule or imperative is precisely the fact that a social norm is conditional. An unconditional imperative might tell us to "trust, no matter what," or "always reciprocate." A social norm instead tells us to trust/reciprocate under various conditions, and to stop trusting/reciprocating if these conditions are not met. In this sense we may think of a norm as subsuming several different strategies that produce the same behavior under the right circumstances. Even more important, we show that the same social norm may be supported by different types of strategy combinations, where the exact polymorphism will depend upon the initial set of basic strategies.

As Brian Skyrms (1996, 1997) has repeatedly argued, a crucial element in the study of norms is an analysis of their emergence. Only a dynamic model allows us to see the history of the emergence of a norm, and the reasons for its change or persistence. In our case, the dynamics are driven by payoff-relevant information, and the use of a simulation lets us see what happens along the path, or how strategies evolve over time. There are advantages to using simulations instead of calculating the limit of the process using a set of difference equations.[8] A simulation eases the process of testing alternative hypotheses that is crucial for understanding how and why different strategies, in different proportions, may converge to a fixed point, or an invariant distribution of strategies. In addition, stochastic

6. If the dynamics were to lead to the survival of a single strategy, we would have a monomorphic population, and the concept of an evolutionarily stable strategy would apply.

7. For a precise definition of social norm, cf. Bicchieri (forthcoming).

8. We use non-linear difference equations (see for example equation (2)) because of discrete time intervals.

difference equation systems are notoriously hard to analyze and the difficulty is compounded—in the case of our model—by the existence of a large strategy space.

It is important to notice that the results obtained in our deterministic simulation are confirmed in the stochastic one. In the latter simulation, the dependence on initial conditions (the initial population proportions) is removed; hence the stable long-run equilibria we observe are a function of the length of the iterated game and the given strategy mix only. In both kinds of simulations, the final stable mix of conditionally cooperative strategies will depend upon the strategies that are currently being played.[9] However, even in environments in which more diverse and complex conditional strategies are present, the final result will be a mix of conditionally cooperative strategies, and unconditional noncooperators will tend to die out, whereas unconditional cooperators will find a niche within the population. We may conclude that, in the long run, norms of trust and reciprocity tend to emerge, provided the initial population contains some conditionally cooperative strategies, irrespective of the specific methods used by these strategies to elicit reciprocity and punish transgressions.[10]

2. The Trust Game. In an interaction, a player can be either in the role of the truster (sender), or in the role of the trustee (receiver). To model the fact that a player has no control over which role he will be cast in, we use an extensive form game (Figure 1) in which Nature moves first. With probability p, Nature assigns to player i the role of sender and to player j the role of receiver, and the reverse with probability $(1-p)$. This game is a version of the 'investment' or 'trust' game studied by Kreps (1990), Bicchieri (1993), and Berg et al. (1995). After the players' roles have been assigned, the sender moves first and must decide whether or not to 'trust' the receiver with her endowment of x dollars. If the sender chooses to trust (invest), the size of her endowment is tripled to $3x$, and the receiver must then decide whether to reciprocate, returning $3x/2$ to the sender and keeping $3x/2$ for himself, or to not reciprocate and keep all $3x$ for himself. The action set for the sender is to trust or not to trust, $a = \{T, nT\}$, and the action set for the receiver is to reciprocate or not reciprocate, $b = \{R, nR\}$.

9. As we explain later on, there are several different definitions of stability for dynamical systems. Whereas in the deterministic case we obtain asymptotically stable states, in the stochastic case we obtain 'stochastically stable' states (see also footnote 28).

10. By 'long run' we refer both to the number of rounds per game, that must be big enough to support cooperation, and to the number of simulations. In our case, with 1000 time-steps there is convergence to a generalized trusting/reciprocating behavior.

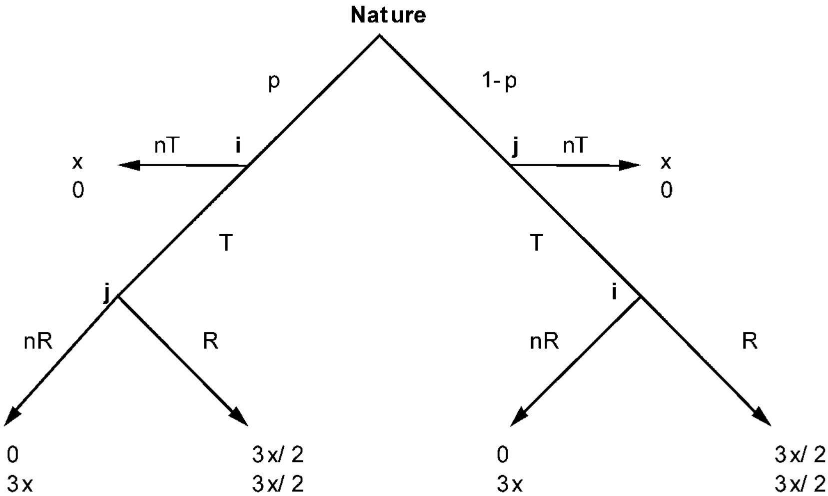


Figure 1.

Once a role has been assigned, each player faces a simple 2x2-payoff matrix, depicted in Figures 2a and 2b. In Figure 2a, the sender is the row player, and clearly has no dominant strategy. In Figure 2b the receiver is the row player, and he has a weakly dominant strategy: nR.

Players, however, have to choose a strategy before knowing which role they will be assigned by nature. In the simplest case, which we examine first, each player has four strategies to choose from: {TR, TnR, nTR, nTnR}. Let us call this set of strategies the *minimal strategy space*. These

a

Sender

	R	nR
T	3x/ 2	0
nT	x	x

b

Receiver

	T	nT
R	3x/ 2	0
nR	3x	0

Figures 2a and 2b.

strategies are role-contingent, in that they dictate how the player should play the game when he finds himself in each role:

TR: when in the sender position, always trust. When in the receiver position, always reciprocate.
TnR: trust as sender, do not reciprocate as receiver.
nTR: do not trust as sender, but always reciprocate as receiver
nTnR: when sender, do not trust. When receiver, do not reciprocate.

Notice that these strategies, while role-contingent, are not conditional on a player's past history of play. In this sense, this set of strategies may be regarded as unconditional, but role-contingent imperatives. For example, TR may be interpreted as the imperative of being unconditionally cooperative, no matter what. Later on we will enlarge the strategy space to include both the minimal strategy space as well as strategies that are conditional on a player's past history of play.

In the remainder of the paper, we shall assume for simplicity that the probability of being cast in either role (sender/receiver) is $p = 1/2$; this will allow us to construct a single, symmetric payoff matrix for the trust game. In the game of Figure 1, then, each player is facing the following 4x4 payoff matrix (Figure 3), which takes the weighted average of playing sender or receiver with probability 1/2. As in Figure 2b, where nR weakly dominated R, here nTnR weakly dominates nTR, and TnR weakly dominates TR.

In what follows we consider a sequence $g= 1, 2, \ldots,$ n, of both one-shot (OS) and repeated (R) games represented by a symmetric payoff matrix **A**, such as that given in Figure 3. In OS games, the game number g, and round number $t = 1, 2, \ldots$ will be synonymous, while in R games, the round number t will start anew, $t = 1, 2, \ldots$ for each new R, "supergame" number g. The main difference between the two environments is in the realization of payoffs; in the OS game, payoffs are realized after a single round, whereas in the R game, the sum of payoffs from all rounds played is realized at the end of each supergame, consisting of a sequence of rounds.

In both environments, each player plays a single pure strategy from the given set of strategies in all rounds of a game. As explained in further detail below, the payoff to using a particular strategy does not depend on the strategy adopted by a player's opponent or sequence of opponents. Rather, we consider how each strategy fares against the population of strategies as a whole—what Maynard Smith calls "playing the field."[11]

If $g \leq n$, then following the completion of each OS or R game, the fitness of each strategy is evaluated. The fitness of a strategy in the OS game is its weighted average payoff against the population of strategies in

11. Cf. Maynard Smith 1982, 23.

	TR	TnR	nTR	nTnR
TR	3x/ 2	3x/ 4	3x/ 4	0
TnR	9x/ 4	3x/ 2	3x/ 4	0
nTR	5x/ 4	5x/ 4	x/ 2	x/ 2
nTnR	4x/ 2	4x/ 2	x/ 2	x/ 2

Figure 3.

the one-round game. The fitness of a strategy in the R game is its weighted average payoff against the population in all rounds of the R game. These fitness values are used to adjust the proportion of the population that is playing each of the pure strategies in the subsequent OS or R game, as explained in further detail below. In the repeated game, we imagine that players discount future earnings by the factor 1-δ per round, where $\delta \in (0, 1)$ can be interpreted as the constant probability that the game ends from one round to the next. Thus, the mean number of rounds in each of the n repeated games is $r = 1/\delta$.[12]

12. If we were to take players to be rational, forward-looking agents, it would make sense to describe r as the *expected* number of rounds. However, since we will only deal with adaptive, or boundedly rational, agents, in all our simulations we will take r to be the *mean* number of rounds per game for a given δ. Also note that our adaptive agents cannot distinguish between a finitely and indefinitely repeated game. However, rather than simulating a sequence of finite games of varying length, we chose to use the mean number of rounds per game in all our simulations.

3. Equilibria in OS or R Games. In the OS game, the unique subgame perfect Nash equilibrium is for receivers to play nR, and therefore for senders to play nT, i.e., in the symmetric game, all players play the strategy nTnR. In the repeated game, it is possible for players to cooperate with each other, and thus do better. Indeed, the *folk theorem* for repeated games states that, for a low enough discounting of future payoffs, there are many equilibria in which the players behave cooperatively towards each other (Fudenberg and Maskin 1986). For example, a trusting-reciprocating behavioral pattern can be supported as an equilibrium if *pairs* of players adopt a "grim-trigger" strategy. Suppose a player in the role of sender initially plays T, and keeps playing T if the receiver plays R, but after the first defection (nR), she switches to nT forever. Similarly, the player in the role of the receiver responds with R to T, but if nT is played once, he plays nR forever after. This grim-trigger strategy can be supported as an equilibrium as long as $\delta \leq 1/2$.[13] Other trigger strategies can be used to support a cooperative, trust/reciprocate equilibrium under similar restrictions on δ. These cooperative equilibria exist also in the population game environment that we examine, where players' strategies are judged by how well they fare against the population of strategies (against the field). In the latter case, the initial proportions of the various strategies in the population may also be a factor (in addition to δ) in whether trust and reciprocity can be sustained (as is shown in our simulations, discussed below).

In sum, as traditional game theory predicts, if players are rational, trust and reciprocity are not sustainable in the one-shot game. However, trust and reciprocity can emerge as stable behaviors when interaction is repeated if players do not discount their future earnings too much, or if they believe with a high enough probability that they are going to meet again.

There are several drawbacks to traditional game-theoretic models. For one, in the repeated game there can be *multiple equilibria,* and no way to predict which one will occur. To play a particular cooperative equilibrium, players must have common priors about all the possible strategies each of them may use, and this fact must be common knowledge. Thus traditional game-theoretic models impose rather heavy informational and computational requirements upon the players.

Furthermore, rationality alone cannot explain how players learn to play a Nash equilibrium, nor can rationality arguments be used to select from among multiple equilibria.[14] Researchers have therefore turned to *bounded*

13. The receiver will choose nR if $3x + (1/\delta - 1)0 > 3x/2(1/\delta)$, where δ is the constant probability that the game ends from one round to the next and the expected number of rounds after the first is $1/\delta - 1$. For him to choose nR, δ must be greater than 1/2.

14. The literature on refinements of Nash equilibrium seeks to solve the equilibrium selection problem by appealing to various rationality arguments (Van Damme 1991). Bic-

rationality approaches to model the adoption and selection of Nash equilibria. Bounded rationality approaches in game theory can be divided into two types, depending on whether the focus is on individual behavior or on population dynamics instead. Individual learning theories, for example, assume some exogenous process for matching players, and describe the manner by which individual players update their beliefs, e.g., the "fictitious play" learning model (see, e.g., Skyrms 1990). Individuals are assumed to play best responses to their most recent beliefs. By contrast, evolutionary theories are inspired by population biology (e.g., Maynard Smith and Price 1973). These theories dispense with the notion of the individual, as well as with best responses/optimization, and use in their place a natural selection, 'survival-of-the-fittest' process together with mutations to model the frequencies with which various strategies are represented in the population over time. We have chosen to pursue an evolutionary learning approach using computer simulations to obtain our main findings.

In the following simulations, players are identified with a strategy, and the relative frequency of a strategy in a population is simply the proportion of players in that population who adopt it. The relative frequency of each strategy in the population at round t is a function of its payoff relative to the population average. In our deterministic evolutionary model, inferior strategies die out, but it is difficult to keep this interpretation whenever a player is identified with a strategy.[15] Players do not necessarily die out. Instead we can suppose that they may simply change their strategies. Thus a more complete model of how a trusting/reciprocating behavioral pattern could emerge as the dominant one would include a description of how adaptive players change their strategies on the basis of previous outcomes.

The main solution concept used in evolutionary game theory is the evolutionarily stable strategy (ESS) introduced by Maynard Smith and Price (1973), or the evolutionarily stable state (ESSt) for population-wide frequencies of strategies (see, e.g., Hofbauer and Sigmund 1998).

---

chieri (1993, ch. 3) discusses some of the reasons why such attempts have had little success.

15. Note that, using the deterministic replicator dynamics, strategies can become extinct in finite time since we are using a finite population with renormalization. The fact that all strategies earn a non-negative payoff is not relevant. The updating procedure associated with the replicator dynamic (our equation 1) is such that strategies increase in the population only if their fitness value is above average. Strategies with below-average fitness values are displaced. A strictly dominated strategy will have below average fitness in every round. Over successive periods, its proportion in the population will steadily decrease, and can indeed become zero in the finite population environments that we consider (see Foster and Young, 1990). Figure 5 illustrates the possibility of extinction in the four-strategy case. To prevent extinction, we need to add noise to the replicator dynamics.

Since we are examining behavior in a population game in which each player plays a pure strategy, we shall adopt the latter concept.[16] Suppose there are N pure strategies for the trust game, with an NxN symmetric payoff matrix $\mathbf{A} = (a_{ij})$ (Figure 3 gives $\mathbf{A}$ for the case where $N = 4$). Each member of the continuum of players initially commits to playing exactly one of the N pure strategies (we do not allow mixtures). Let p be the $N \times 1$ vector denoting the population-wide proportion of each of the N strategies (player types) in the population. Let

$$f_i(p) = \sum_j a_{ij} p_j = A_i p$$

denote the fitness of strategy i. The population-wide weighted average fitness value is $p^T Ap$. We say that $\hat{p}$ is an *evolutionarily stable state* (ESSt) if, for any p

$$\hat{p}^T A\hat{p} \geq p^T A\hat{p}.$$

And if $p \neq \hat{p}$ and $p^T A\hat{p} = \hat{p}^T A\hat{p}$, then

$$\hat{p}^T Ap > p^T Ap$$

The first inequality is just the definition of a Nash equilibrium. The second inequality is a further refinement that guarantees that $\hat{p}$ is not invadable; that is, $\hat{p}$ fares better against p than p fares against itself.

The definition of an ESSt does not refer to a specific dynamic, but biologists and evolutionary game theorists frequently use a *replicator* dynamic, which in its deterministic form can be written as:

$$p_i(t+1) = \frac{p_i(t) A_i p(t)}{p^T(t) Ap(t)}, \tag{1}$$

where $p(t)$ denotes the population-wide proportion as of time t. Hence, strategies with above average fitness see their proportions increase, and those with below average fitness see their proportions decrease.[17]

16. Of course, a population can play an evolutionarily stable strategy. If we allow *mixtures* then, as Maynard Smith showed, a certain mixed strategy, e.g., where *all* players play hawk (dove) according to a certain fixed probability, can be an ESS. We do not allow mixtures, as we adopt the biological convention that each player is a particular phenotype, and can be thought of as having a single, pure strategy (either a fixed or a conditional rule). One could further argue that mixtures across such pure strategies are difficult to interpret. With only pure strategies, the relevant solution concept is the "evolutionarily stable polymorphic state," the limit or rest point of an evolutionary process. Cf. Maynard Smith (1982, 11, and also Appendix D).

17. It should be noted that in the rest of the paper we use a discrete replicator dynamics (since we assume the population is finite).

It is well known that (ESSt) are asymptotically stable fixed points of this replicator dynamic, though the converse of this statement need not be true (see, e.g., Samuelson 1997).[18] A similar relationship holds between the replicator dynamic and Nash equilibria: if $\hat{p}$ is a Nash equilibrium of the symmetric game A, then $\hat{p}$ is a stationary state of the replicator dynamic.

4. The Evolutionary Model. We use both a deterministic and stochastic discrete-time replicator dynamic to characterize the distribution of strategies in the population over time. The system has an evolutionary flavor in the sense that each strategy's share in the population increases or decreases with increases or decreases in that strategy's payoff performance relative to the population average payoff. Furthermore, in the stochastic version of the model, *mutations* in the proportions of the various strategies in the population insure that no strategy becomes extinct.[19] This simple, dynamical model is based on one used in population biology and was imported into game theory by Foster and Young (1990; Young and Foster 1991), who introduced the idea of a stochastically stable equilibrium.

4.1. Deterministic Version. The deterministic version is as described above. There are N strategies for the trust game and $p(t)$ is the $N \times 1$ vector denoting the proportion of each of the N strategies in the population at time t. For example, in the simplest population we examine, where $N = 4$,

18. A similar point was made by Maynard Smith (1982, Appendix D). Suppose P is an evolutionarily stable mixed strategy used by an *individual* player who can use all the pure strategies in the given strategy set (e.g., in the hawk-dove game, the mixture might be hawk with probability .60 and dove with probability .40). Now suppose we rule out such mixed strategies, and let p be the frequency of pure strategy types (or phenotypes) in a polymorphic population. What Maynard Smith shows is that if there are just two pure strategies, and P is an ESS mixed strategy, then if players play only pure strategies we will have P = p, that is, the population of pure strategy players will converge to a polymorphism where 60% are hawks and 40% are doves, the population analogue of the mixed ESS. More generally, if there are more than two pure strategies, and P is an ESS mixed strategy, then the corresponding polymorphism p = P will be stable. However, when there are more than two strategies a stable polymorphism p *does not imply* that the corresponding mixed strategy P is an ESS, as we note in the paper. The reason is that stable polymorphisms (in pure strategies) might be invaded by mixed strategies.

19. In most versions of the deterministic model, if a strategy does not survive the iterated elimination of strictly dominated strategies, then that strategy also does not survive under most versions of the replicator dynamics. This theorem is proved in Samuelson and Zhang (1992), but the logic is intuitive. Suppose there are just two strategies, and one strictly dominates the other. The dominant strategy will have a higher relative fitness value and so will increase its proportion in the population, while the dominated strategy will decrease its proportion and die out. This generalizes to more than two strategies.

$p(t) = (p_{TR}(t), p_{TnR}(t), p_{nTR}(t), p_{nTnR}(t))$ and there is no population growth, i.e., $\sum_i p_i(t) = 1 \;\; \forall t$. The $\mathrm{N} \times \mathrm{N}$ matrix of payoff values for either the one shot or the repeated game is denoted by **A**. This matrix summarizes the expected payoff earned by each strategy when matched against each of the other strategies in the population, including itself. The deterministic version of the evolutionary model has the proportion of the population using strategy *i* evolve according to the simple replicator dynamic given in (1).

4.2. Stochastic Version. Foster and Young (1990) pointed out that the deterministic system (1) allows some strategies to become extinct, in the sense that $\mathrm{p_i(t)} = 0$ for some i, t. The possibility of extinction runs counter to the (biological) notion that populations are subject to invasion and strategies (species) that are near extinction may thrive once again when environmental conditions change in their favor. Furthermore, there is the possibility of new strategies (species) as well, however we do not consider this possibility here. To prevent extinction, we can add mutation to the model in several ways. For example, we can perturb the payoff matrix **A** slightly each period as in Fudenberg and Harris 1992, or we can add noise to the deterministic updating of the proportion vector *p*(t) as in Foster and Young 1990, or we can do both. Perturbing the payoff matrix can be interpreted as uncertainty concerning the expected payoffs, while perturbations to the proportion vector can be interpreted as persistent experimentation (or non-extinction).

We propose, as our stochastic model, the latter type of mutation. Specifically, let the proportions now evolve according to:

$$p_i(t+1) = \frac{p_i(t)A_i p(t)}{p(t)^{\mathrm{T}} A p(t)} + s\,|\,\varepsilon_i(t+1)\,|, \tag{2}$$

where $\varepsilon_i(t+1)$ is a draw from a standard normal distribution, and s is a tuning parameter. The algorithm we used is implemented as follows. In each period, we calculate the proportions according to (2). We then *rebalance* these proportions, by dividing each $p_i(t+1)$ by the sum $\sum_{i=1}^{N} p_i\ (t+1)$. Our interest is in the evolution of the proportion vector $\vec{p}$ (*t*) over time for "small" values of *s*. Foster and Young showed that the behavior of the stochastic system can be quite different from the behavior of the deterministic system. In particular, the stochastic system removes the possibility of absorbing states at the boundary of the $N-1$ dimensional simplex that characterizes the distribution of strategies in the population, and under certain conditions, can result in a unique, "stochastically stable" equilibrium proportion vector.

Matrix A	TR	TnR	nTR	nTn R
TR	1.5	0.75	0.75	0
TnR	2.25	1.5	0.75	0
nTR	1.25	1.25	0.5	0.5
nTn R	2	2	0.5	0.5

Figure 4.

Because we are interested in examining the emergence of a behavioral regularity in a relatively large strategy space (later on we will examine up to 16 strategies), analytic results are difficult to achieve. We therefore make use of numerical simulations as described in the next section.

5. Simulations.[20]

5.1. One-Shot Game. In the one-shot game, **A** is the 4x4 symmetric matrix depicted in Figure 3. In the simulations, we set $x = 1$, so the actual matrix used is given in Figure 4. All of the simulation experiments reported in this section were conducted using the simple deterministic replicator dynamic (1).[21] The initial proportions of strategies were varied according to what we thought were interesting initial conditions, and the simulations were carried out for a sufficiently large number of periods (1,000 iterations for each initial condition) to insure that the limiting, stationary proportions of the replicator dynamic had obtained. Each round in a simulation corresponds

20. The *Mathematica* code used to carry out our simulations is available at http://www.pitt.edu/~jduffy/trust/.

21. We ran six simulations that differed in the initial proportion of players playing the different strategies.

TABLE 1.

Initial Condition #	Strategy	Initial Frequency	Final Frequency*
1	TR	1	1
	TnR	0	0
	nTR	0	0
	nTnR	0	0
2	TR	0	0
	TnR	1	1
	nTR	0	0
	nTnR	0	0
3	TR	0	0
	TnR	0	0
	nTR	1	1
	nTnR	0	0
4	TR	0	0
	TnR	0	0
	nTR	0	0
	nTnR	1	1
5	TR	0.99	0
	TnR	0.01	1
	nTR	0	0
	nTnR	0	0
6	TR	0.99	1
	TnR	0	0
	nTR	0.01	0
	nTnR	0	0
7	TR	0.99	0
	TnR	0	0
	nTR	0	0
	nTnR	0.01	1
8	TR	0.01	0
	TnR	0.99	1
	nTR	0	0
	nTnR	0	0
9	TR	0	0
	TnR	0.99	1
	nTR	0.01	0
	nTnR	0	0
10	TR	0	0
	TnR	0.99	0
	nTR	0	0
	nTnR	0.01	1
11	TR	0.01	1
	TnR	0	0
	nTR	0.99	0
	nTnR	0	0
12	TR	0	0
	TnR	0.01	1
	nTR	0.99	0
	nTnR	0	0
13	TR	0	0
	TnR	0	0
	nTR	0.99	0.99
	nTnR	0.01	0.01
14	TR	0.01	0
	TnR	0	0
	nTR	0	0
	nTnR	0.99	1
15	TR	0	0
	TnR	0.01	0
	nTR	0	0
	nTnR	0.99	1
16	TR	0	0
	TnR	0	0
	nTR	0.01	0.01
	nTnR	0.99	0.99
17	TR	0.97	0
	TnR	0.01	0
	nTR	0.01	0.000577428
	nTnR	0.01	0.999423
18	TR	0.01	0
	TnR	0.97	0
	nTR	0.01	0.000577428
	nTnR	0.01	0.999423
19	TR	0.01	0
	TnR	0.01	0
	nTR	0.97	0.0673118
	nTnR	0.01	0.932688
20	TR	0.01	0
	TnR	0.01	0
	nTR	0.01	0.00989659
	nTnR	0.97	0.990103
21	TR	0.25	0
	TnR	0.25	0
	nTR	0.25	0.121004
	nTnR	0.25	0.878996
22	TR	0.25	0
	TnR	0.25	0
	nTR	0.01	0.00666536
	nTnR	0.49	0.993335
23	TR	0.25	0
	TnR	0.25	0
	nTR	0.49	0.0303428
	nTnR	0.01	0.969657
24	TR	0.5	1
	TnR	0	0
	nTR	0.5	0
	nTnR	0	0
25	TR	0.48	0
	TnR	0.02	1
	nTR	0.48	0
	nTnR	0	0
26	TR	0.48	0
	TnR	0	0
	nTR	0.48	0.0482243
	nTnR	0.02	0.951776
27	TR	0.48	0
	TnR	0.01	0
	nTR	0.48	0.0307279
	nTnR	0.01	0.969272

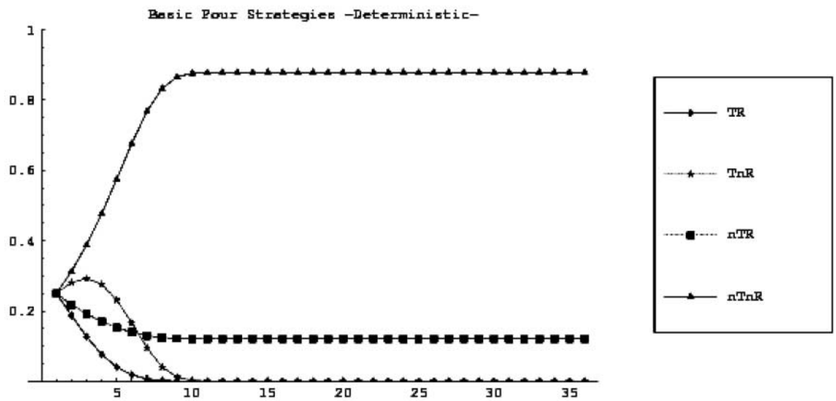


Figure 5.

to the play of a single one-shot game. Our results consider the proportion of the various strategies observed in the population at the end of 1,000 iterations.

The main results can be summarized as follows. First, as long as there is initially some positive proportion of the population playing nR, the proportion of players playing R (either as TR or nTR) will disappear.[22] If the initial proportion playing nR is zero, then T will come to dominate nT, as long as there is some positive proportion of players initially playing T.

Table 1 shows 27 initial frequencies (conditions) for the basic four strategy version of our game, and what happens to these frequencies following 1000 iterations of the deterministic replicator dynamic. Note that in a simulation where the initial population has any positive proportion of the nTnR strategy, the final result will be a population dominated by nTnR players (but see condition 13 below for an exception). For example, if the initial vector $p = (.25, .25, .25, .25)$, which corresponds to condition 21 in Table 1, the limiting proportion vector as illustrated in Figure 5 is: .0 TR, .0 TnR, 0.121 nTR and 0.879 nTnR. This is the limiting vector because the boundary is an absorbing state. As Figure 5 illustrates, the convergence in the deterministic case happens after about 10 iterations, so 1000 iterations is plenty.

Similarly, if the initial population is composed only of nTR players and nTnR players, then both strategies will survive, because if no players trust, then the receiver's strategy does not matter. For example, if $p\,(0) = (.0, .0, .99, .01)$, as in condition 13 of Table 1, the system does not move away at all from this initial condition, where we have .99 nTR and .01 nTnR. This can happen because, if almost everyone is playing nTR, there is never

22. In the deterministic case, extinction is possible for the reasons outlined in footnote 15.

an opportunity for nR to spread, as receivers never get to make any choice. In general, many initial conditions will end up on the boundary between nTR and nTnR, though far closer to nTnR. These anomalous results disappear once we add a stochastic element to the population updating procedure.

The conclusion of our simulation exercise for the one-shot game is that we get stable polymorphic equilibria. However, any population vector composed entirely of non-trusting strategies will not change over time, hence we may consider this vector an evolutionarily stable state.

5.2. Repeated Game. In the repeated game setting, each game consists of a number of rounds, r. In each round, each player type (defined by its strategy) plays its strategy against all player types (including itself) according to their proportions in the population, yielding a certain population-weighted payoff for each strategy for each round. The strategy proportions do not change until the end of the r-rounds. At the end of each repeated game, the proportions of the various types of players are updated according to the replicator dynamic given in equation (1), where the fitness of strategy i is based on its performance *over all* r rounds of the game.

Recall that the value of $r = 1/\delta$, where δ can be interpreted as the constant probability that the game ends from one round to the next (alternatively, $1-\delta$ is the constant discount factor for payoffs). Thus, r increases as δ decreases. In all our repeated game simulations, we varied δ from 1 to .0333, so that r varied from 1 (one-shot game) to 30 rounds of play.[23]

In the repeated game setting we introduce conditional strategies, since repetition allows for more complex behavior. In what follows, we consider a minimum number of conditional strategies to better understand their individual contribution to the establishment of a cooperative behavioral pattern. There are four sending and four receiving strategies:

Sender

1. Always trust (T).
2. Never trust (nT).
3. Grim trigger (G) – Trust until you are not reciprocated and then do not trust for the rest of the interaction.

23. Alternatively, we could have considered a true indefinitely repeated game, in which case the *mean* number of rounds in each game would be given by $r = 1/\delta$. In this case, there would be considerable variation from this mean number of rounds across games. Indeed, there would also be the (slight) possibility that the game would continue indefinitely, so that some kind of truncation or upper bound on the number of rounds played would be necessary. As the strategies adopted by our agents were not forward-looking (so they could not use backward induction) we chose to forego the complications associated with an indefinitely repeated game, and we simply varied the finite length of the repeated game, r, as described above.

4. Hopeful (H) – Trust, and then if you are not reciprocated, retaliate by not trusting in the next round. The round after that, trust again, as a sign of your willingness to trust. If you are still not reciprocated, do not trust for the rest of the interaction. If you are reciprocated, return to the beginning of this strategy.

The Hopeful strategy is one of many possible retaliatory strategies. When playing Hopeful, a player stops signaling her willingness to trust after she gets a non-reciprocating response for two rounds, but one can build other strategies in which a player keeps signaling her willingness to trust in the face of several defections. In fact, one could make "hopefulness" a function of the stopping probability, i.e., the lower the probability the game will end soon, the more likely it is that a player "tries again" to signal her willingness to trust. As we explain in footnote 25, we tried other, more complicated 'hopeful' strategies that employ more rounds of punishment. However, we found that these alternative strategies always did marginally worse that the standard Hopeful one.

Receiver

1. Always reciprocate (R).
2. Never reciprocate (nR).
3. Grim trigger (G) – Reciprocate until you are not trusted, then do not reciprocate for the rest of the interaction.
4. Selfish (S) – Start by not reciprocating if you are trusted. If you are not trusted in one round, switch to reciprocating in future rounds in which you are trusted.[24]

The Selfish strategy is one in which the receiver "tests" the sender. If the sender retaliates, the receiver switches to cooperative behavior.

Given the four sender and receiver strategies, there are 16 possible combinations of strategies for players in the repeated game. These are: TR, TnR, TG, TS, nTR, nTnR, nTG, nTS, GR, GnR, GG, GS, HR, HnR,

24. Another strategy we considered in the present simulation is a Hopeful strategy for the receiver. It would start by reciprocating if trusted, and if not trusted in one round, it would punish the sender who subsequently trusts by not reciprocating once, but switching to reciprocating again if trusted in subsequent rounds. This strategy, when combined with our current strategies, would never have a chance to show that it was different. To trigger this strategy's retaliation, a trustor would have at some point not to trust, and then trust again. Two of our strategies either trust unconditionally, or do not trust unconditionally. The Hopeful sending strategy would not trigger retaliatory behavior, as the Hopeful responder would always reciprocate. We would have to introduce a strategy that delays trusting, perhaps starting off with not trusting before attempting to trust, but as no information can be gained from not trusting initially, this strategy will always lose out to one that trusts. In essence, if you are going to trust at all, there's no reason to not do it immediately.

a

Sender	R	nR	G	S
T	1.5r	0	1.5r	0
nT	r	r	r	r
G	1.5r	r-1	1.5r	r-1
H	1.5r	HnR(r)	1.5r	HS(r)

b

Receiver	T	nT	G	H
R	1.5r	0	1.5r	1.5r
nR	3r	0	3	nRH(r)
G	1.5r	0	1.5r	1.5r
S	3r	0	3	SH(r)

Figures 6a and 6b.

HG, and HS. Figures 6a and 6b represent, respectively, the 4x4 payoff matrices for sender and receiver. Each element of the matrix is the cumulative payoff obtained at the end of *r* rounds of play between the column strategy and the row strategy. These matrices are expanded into 16x16 matrices, as before, and then averaged to represent the equal probability of being cast in the role of sender or receiver. Some payoffs have been expressed as piecewise functions. For example, in the sender's matrix, we have that, HnR (1) = 0, HnR (2) = 1, HnR (x) = x−2, HS (1) = 0, and HS (x) = 1.5x−2. In the receiver's matrix, the values would be nRH (1) = 3, nRH (2) = 3, nRH (x) = 6, SH (1) = 3, and SH (x) = 1.5x.

We use the same deterministic replicator dynamic (1) as before. However, the payoff matrix **A** is now the larger 16x16 payoff matrix.[25] In the simulation exercises reported below, we varied *r*, the number of rounds played (equivalently, the inverse of the discount factor) from 1 to 30. We also experimented with different initial proportions of strategies in the population. For each set of initial proportions and value of *r*, we simulated the deterministic replicator dynamic for a sufficiently long time (1000 periods) to ensure that the limiting stationary proportions obtained.

6. Results.

6.1. The Deterministic Model. The results of our simulation exercises involving the deterministic replicator dynamic in the repeated game setting depend on the value of *r* and the initial proportions of strategies in the population. These results can be summarized as follows. In the case where there are equal initial proportions of the 16 strategies, the long-run outcome is a function of the discount factor, or the number of rounds played in each interaction, *r* (see Figure 7). On the horizontal axis of Figure 7 (and subsequent figures) is the number of rounds per game. The strategy proportions, represented as bars, are average proportions over 1,000 simulated games, always starting with the same initial condition (equal proportions) but different game lengths. It is interesting to note that several contingent strategies support cooperative behavior, but none of them is an ESS.

25. As mentioned before, we tried to expand the set of "hopeful" strategies. We wondered whether a round of punishment was optimal, and whether more rounds of punishment would be any better. We created two new strategies, termed H2 and HII. The H2 strategy punishes for two rounds before trying to trust again. But if not reciprocated again, it will not trust anymore. The HII strategy will trust initially, and if not reciprocated it punishes by not trusting in the next round. As distinct from the Hopeful strategy, HII will give another chance to the receiver by trusting again and, if not reciprocated again, will again punish for a single round. The third time, it will punish forever the receiver that violates her third attempt to trust. The piecewise cumulative payoff functions are found in an expanded version of this paper at http://www.pitt.edu/~jduffy/trust/. Over many simulation runs, with different mixes, we found that these two strategies always did marginally (<.001) worse than the standard one-try one-round Hopeful strategy. On the basis of this result, we are only including the standard Hopeful strategy in our simulations.

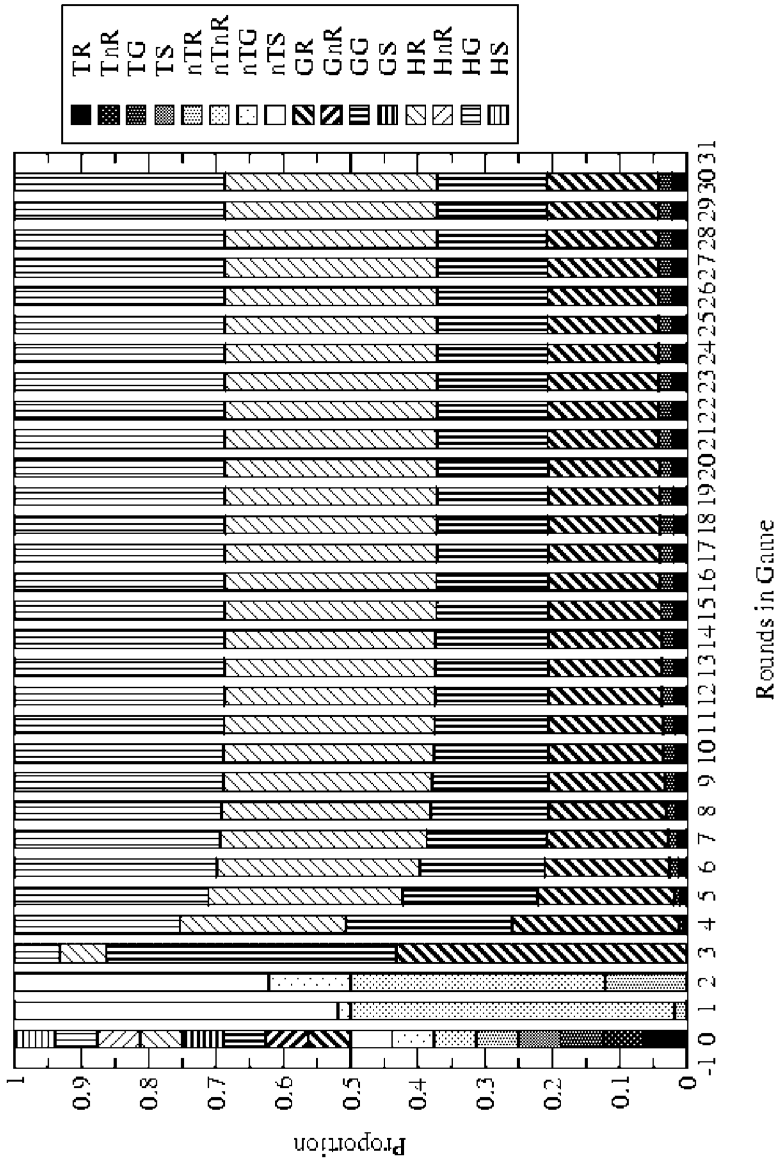

TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS
Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Rounds in Game
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Figure 7.

Notice that when $r = 1$, regardless of the initial population proportions, the long-run outcome is mainly composed of nTnR and nTS in equal proportions, with a small number of players playing nTR and nTG. Exceptions occur when nTnR initially dominates the population, as Figure 10 illustrates. In this case, nTnR remains the most prevalent strategy. This result is not too surprising, since when players in our simulation play a one-shot game 1000 times, non-trusting, non-reciprocating strategies dominate as in the simple one-shot game. In short, impersonal trust does not emerge when the length of interactions is short. These conditions are generally true in the case of $r = 2$ as well, but with a larger proportion of players playing the reciprocating strategies R and G, and fewer playing nR and S.

However, when $r = 3$, we observe a dramatic shift in favor of players playing GG and GR. It is at this game length that the simple learning strategy embedded in the Grim strategy shows some strength. We also observe a small number of players playing HR and HG. For $r > 3$, we increasingly observe players playing HR and HG, while the proportion of players playing GG and GR declines. In most cases, the limiting distribution is a mixture of strategies, with about 60% of the players split between HR and HG, 30% split between GR and GG, and 10% playing TR and TG. In short, the only strategies that survive as r grows large are the contingent strategies that favor trust and reciprocation. As the number of rounds increases, a population composed only of these strategies is stable, as all players trust and reciprocate. In addition, a population mainly composed of smart trusters can support a small number of unconditional, "dumb" trusters. It is important to stress that non-contingent strategies such as TR can only survive because of the presence of contingent cooperative strategies.

6.2. Some Exceptions. The results above were for the case of a uniform initial distribution of the 16 strategies in the population. In a population initially dominated by players playing one of the four non-contingent T strategies (TR, TnR, TG, or TS), as r gets large, the long-run outcome is made up of 60% HS and 40% HR/HG. For example, in Figure 8, the population is initially composed of 85% TR and equal proportions of the other 15 strategies. It appears that the reciprocating strategy that gains the most from the existence of a large initial proportion of unconditional trusters is the Selfish strategy. It gains from unconditional trusters by not reciprocating, but learns to reciprocate when faced with players that do retaliate.

A second exception arises when one of the grim or hopeful sender strategies (GR, GG, HR, HG, and HS) makes up a large proportion of the initial population. For example, Figure 9 illustrates the case where 85% of the initial population plays GR, though the same results would hold if any of the other strategies GG, HR, HG, or HS initially dominated to the same degree. In the one-shot, $r = 1$ game, the result is a distribution of non-trusting

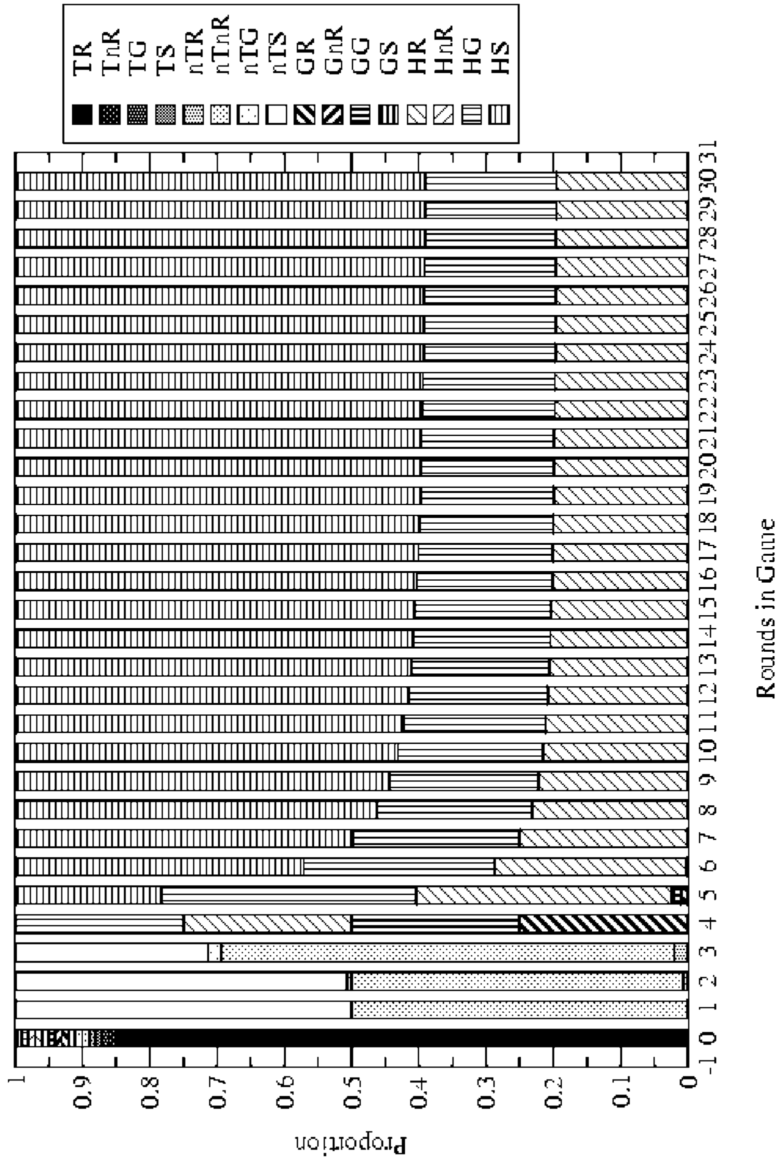
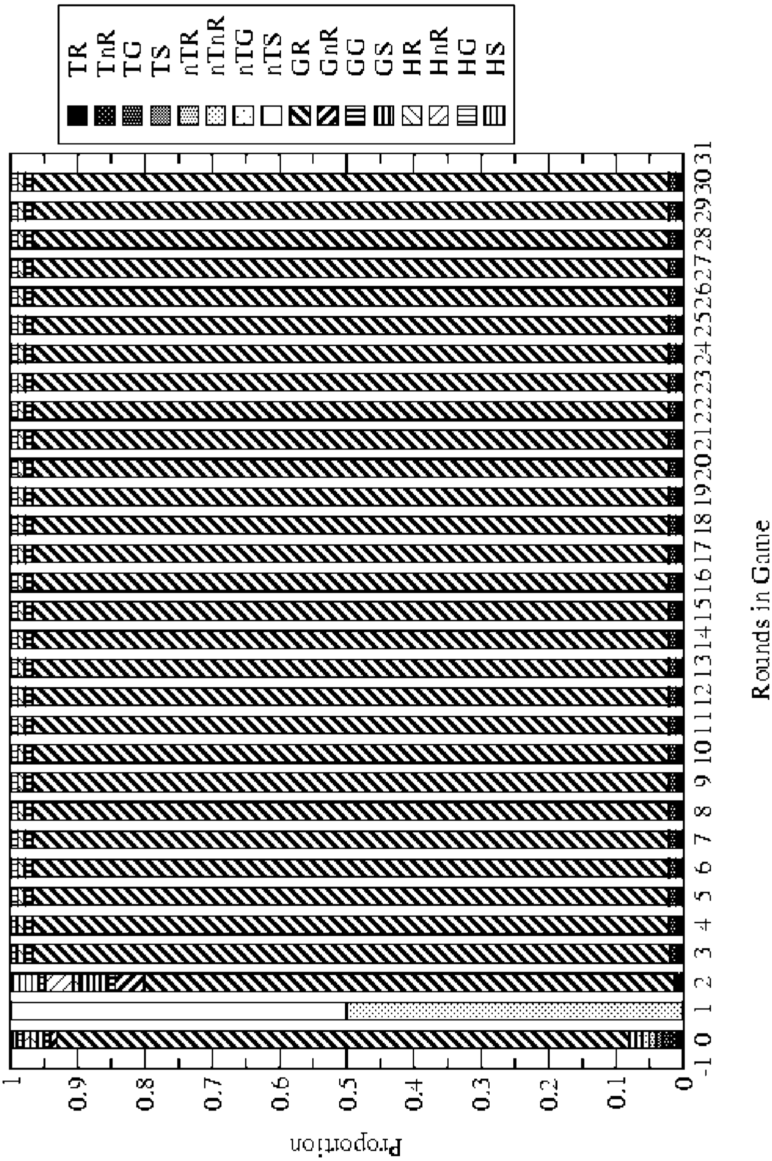

TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS
Proportion
Rounds in Game

Figure 8.



Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Rounds in Game
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS

Figure 9.

strategies. But for $r > 1$, the dominant initial strategy (GR) remains dominant. When these five strategies are played against each other, HS dominates when $r = 1$, remains stable when $r = 2$, and disappears entirely for larger values of r. When nTnR is initially the most frequent strategy, we still observe a shift to the mix of strategies outcome described earlier, but not until $r = 10$ (see Figure 10, and compare it with Figure 7). nTnR, the fittest strategy in the one-shot game, remains strong for low values of r, but once r equals or exceeds 10, this strategy becomes extinct based on the distribution of results from 1000 r-round games. Notice that in Figures 7–10, it appears as though $r = 2$ or 3 is usually the indifference point, or tipping point. Before that, non-trusters dominate, and after that, conditional trusters dominate the population.

6.3. The Stochastic Model. The stochastic version of the simulation models the evolution of strategies using the stochastic replicator dynamic, equation (2). Computation of the strategy proportions at time $t+1$ at first proceeds according to the standard deterministic population equation. The new equation then adds a noise term, composed of the absolute value of a draw from a standard normal distribution, multiplied by the tuning parameter s. This has the effect of adding a small, individual, random positive value to each population proportion. After each population proportion has been computed, the values are renormalized, dividing each one by the sum of the new proportions. This ensures that all the values continue to sum to one, and as some values have had larger random modifications than others, the overall effect is that some strategies will gain from the random element, and some will lose. Another consequence of the stochastic model is that all strategies will survive in the population in some measure, bounded below by the mutation rate.

We begin by examining the behavior of the original four unconditional strategies under this new stochastic model. For all the simulations below, we are using a tuning parameter $s = .01$, so as not to introduce large amounts of randomness into the results.

6.4. Unconditional Strategies. As none of these strategies will perform any differently in a one-shot game as compared to an iterated game, the value of r will not make a difference in the final outcome. Regardless of the initial conditions, we see a convergence (with small fluctuations) to a population vector with mean values: TR = .01, TnR = .01, nTR = .12, nTnR = .86.[26] An illustration with two different initial proportions is provided in Figures 11a and 11b. These results are qualitatively similar to

26. These proportions represent a stochastically stable state of the population, given our choice of noise.

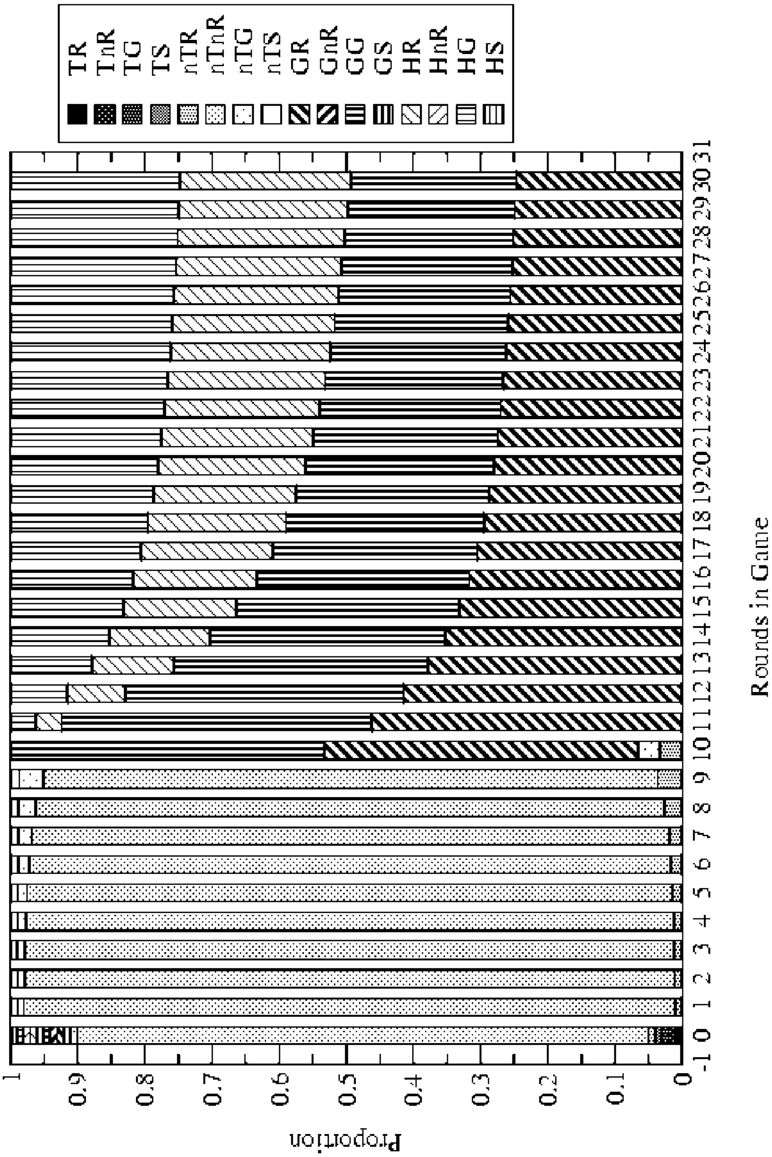
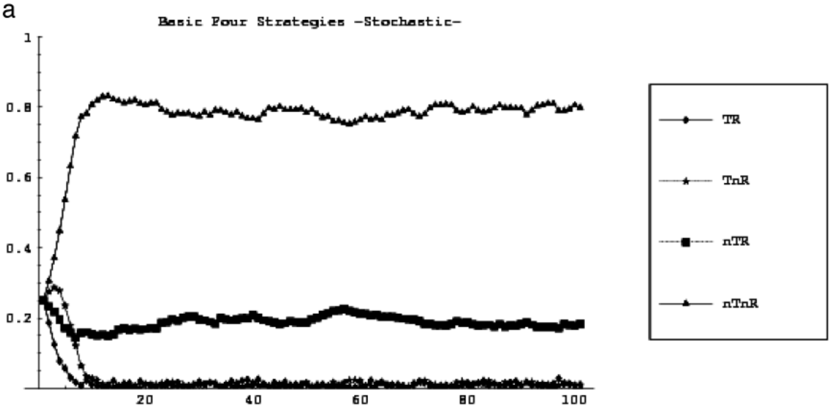
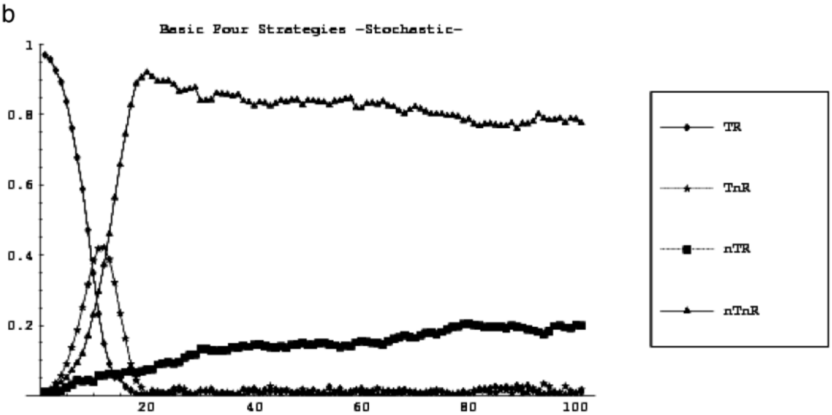

TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS
Proportion
Rounds in Game

Figure 10.

a



b



Figures 11a and 11b.

the deterministic case (compare Figures 11a and 11b to Figure 5). Again, note that we report the average proportions following 1000 games with r-rounds each, where fitness is evaluated at the end of each game. The main difference is the greater volatility in the proportions over time due to the stochastic replicator dynamic. Also, two strategies, TR and TnR are kept from becoming extinct in the stochastic model by the presence of the error term in equation (2)

6.5. Conditional Strategies. When conditional strategies are considered, for small values of r (i.e., $r < 3$), the unconditional non-trusters and non-reciprocators dominate the population, as we would expect. However, the results are quite different, as r gets large. Here we report results from an experiment where we add conditional strategies, one at a time, into a population with unconditional strategies. As before, we have removed all dependence on initial conditions, and we observe convergence to a stochastically stable mix of strategies based on the initial mix and on *r*.[27] The first case we consider consists of five strategies, the basic four unconditional strategies B4={TR, TnR, nTR, nTnR} and the conditional strategy GG. This case is illustrated in Figure 12. For $r > 3$, the mean proportions (over all *r*) after 1000 periods are: TR = .23, TnR = .04, nTR =.04, nTnR = .02, GG = .66. The simple addition of a conditional learning Grim/Grim strategy to the mix completely alters the final equilibrium for large *r*. The GG strategy dominates the population, and the trusting TR strategy is allowed a second place, due to the existence of the conditional GG strategy. The proportions of the other 3 strategies are due to the stochastic replicator dynamic, which prevents these strategies from dying out. We observe a similar finding when we consider the basic four strategies, B4, either combined with GR or with GS. In the latter case, however, we see a new mean proportion mix: TR = .08, TnR = .03, nTR = .50, nTnR = .04, GS = .34. It seems that the selfish strategy, which feeds on unconditional trusters, allows the non-trusting strategy to gain. In addition, the nTR strategy gains from its own unconditional responses to the other trusting strategies in the population.

Consider next the case where the initial strategies are B4 (the basic four) together with GG and HR. In this case, GG and HR share 66% of the population, but GG does a bit better, most likely due to the absence of the Selfish responding strategy, from which the Hopeful trusting strategy can benefit. However, with B4+GG+GS, GG comes to dominate again and reduces GS to less than 5% of the population. The Selfish strategy only seems to work whenever it is possible to take advantage of unconditional trusters.

The next case to consider is an initial population made up of nTnR, GR, GS, HR, and HS. In this case the final mix of strategies in the population, for large *r*, is nTnR = .01, GR = .27, GS = .04, HR = .61, HS = .05. We see that the reciprocating strategy does better across the board than the selfish strategy, as most of the conditional trusting strategies end up punishing the selfish strategy. The uncooperative strategy nTnR does not fare well at all.

Finally, we introduced all 16 strategies together in the stochastic model. The result, illustrated in Figure 13a, is a final mix dominated by HR and HG in roughly equal proportions of 20% each. GR/GG come in second place, at about 10% each. Next we see GS/HS at about 6% each. Most other strategies stay around 4–5%, with non-reciprocating strategies doing worse at around 2%, and finally nTnR doing the worst at about 1%. However these last few strategies survive only due to the stochastic nature of the replicator dynamic, which does not let them disappear. In Figure 13b, we observe that

27. In the limit, as the noise goes to zero, we would have convergence to a unique equilibrium.

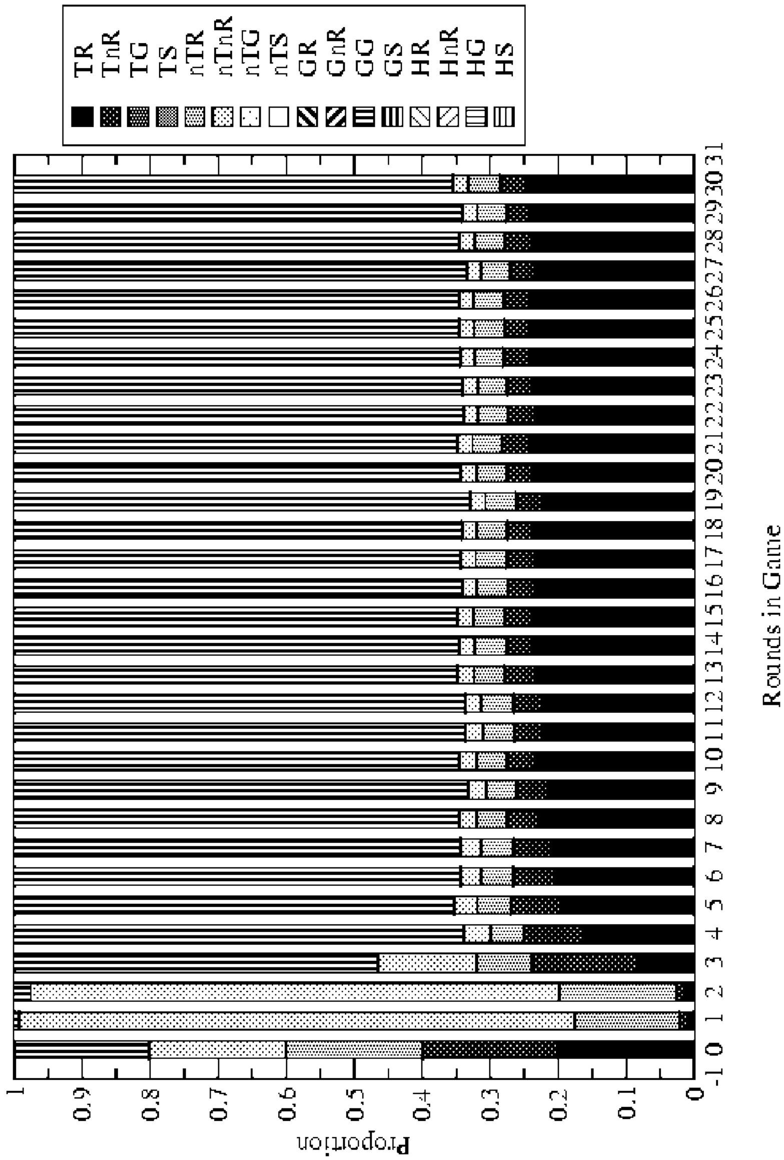

TR
T*n*R
TG
TS
*n*TR
*n*T*n*R
*n*TG
*n*TS
GR
G*n*R
GG
GS
HR
H*n*R
HG
HS
Proportion
Rounds in Game

Figure 12.

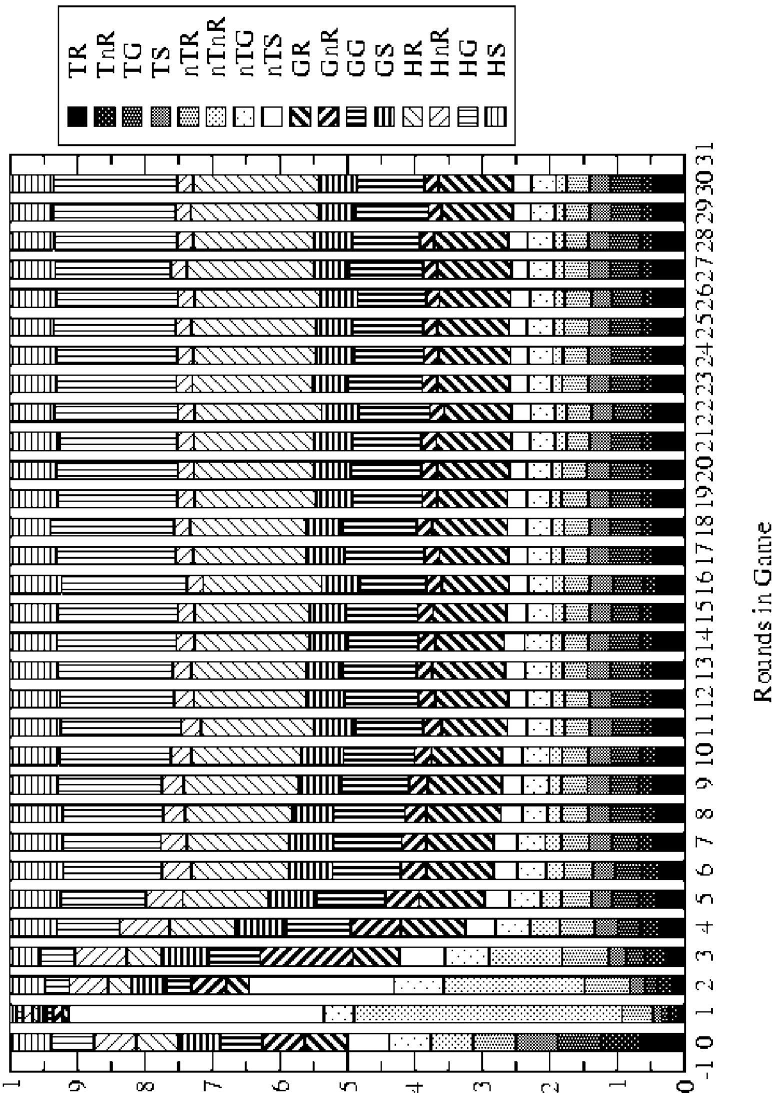

TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS
Proportion
Rounds in Game

Figure 13a.

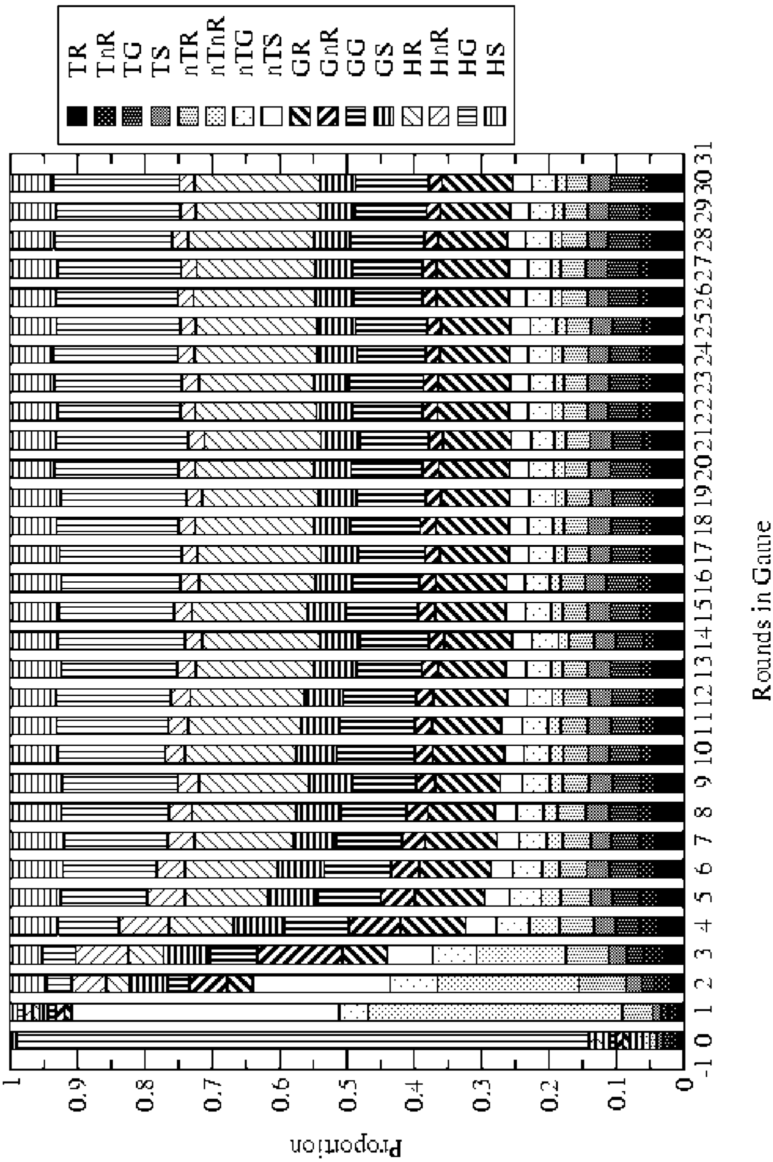

TR
TnR
TG
TS
nTR
nTnR
nTG
nTS
GR
GnR
GG
GS
HR
HnR
HG
HS
Proportion
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Rounds in Game

Figure 13b.

this same distribution continues to hold if one of the two the most successful strategies from the previous experiment, HG, dominates the initial population of strategies.

The most important result we obtain is that regardless of the initial population proportions, in the long run we converge to a stationary equilibrium consisting of a polymorphic population of strategies.[28] Each value of r may result in a different equilibrium outcome but the equilibrium is stable, and is a function only of the initial strategy mix and the value of r, the length of the iterated game. As in the case of the deterministic simulation, as r gets large, conditional strategies come to dominate the population. The unconditional trusting/reciprocating strategies continue to exist, though in small numbers, due to the presence of large numbers of conditionally "nice" strategies. It thus seems that the very existence of a rigid rule that demands unconditional cooperation depends upon the presence of conditionally cooperative practices. As r gets large, non-trusting/non-reciprocating strategies invariably recede to a mere subsistence level. The tuning parameter s represents the weight cast on the stochastic term. Any positive value of s will lead, irrespective of the initial proportions of strategies, to the same equilibrium mix of strategies. For smaller values of s, the system will take more time to converge to an equilibrium mix of strategies, and larger values of s will create more fluctuations in the actual proportions. It will still be the case, however, that the strategies that are dominant in an environment in which there is low variability remain dominant in a high-variability environment.

7. Conclusions. Our goal in this paper was to explicate the development of a social norm of trust and reciprocation. We show that when impersonal trust/reciprocation becomes the dominant observed behavior, it is the outcome of the interaction of several different strategies. It would therefore be a mistake to identify a social norm of trust and reciprocation with a particular strategy, since such a norm is supported by several different strategies. Moreover, such strategies are conditional ones. In fact, many pure strategies support trust/reciprocating behavior, but none of them is an *unconditional* rule telling a player to trust/reciprocate no matter what. Rather, generalized, impersonal trust only develops as a consequence of the interaction of several *conditional* strategies that differ in the severity with

28. There are several different definitions of stationarity for dynamical systems. One is asymptotic stability, wherein a dynamical system converges to a fixed point. But a stationary state need not be asymptotically stable; it can just be Lyapunov stable, or in our case "stochastically stable," by which we mean that after some period t, we reach a stationary distribution of strategies, having some mean p and support (or bounds) $[p - \varepsilon, p + \varepsilon]$, so that in every period after t, the system never leaves those bounds.

which they punish transgressions, as well as in their willingness to give other players further chances. Their differences, however, cancel out in our evolutionary model.

For impersonal trust to emerge, interactions between players must go on for extended periods of time. When interactions are only one-shot, players who take what they can and leave do the best. When interactions are repeated, in the presence of conditionally cooperative strategies, anonymous, impersonal trusting/reciprocating behavior can emerge and dominate in this evolutionary model. The resulting stable behavioral pattern (or norm) is thus supported by a polymorphic population of strategies. Regardless of the proportions in which our strategies initially appear in the population, a resulting evolutionarily stable state is reached in which trusting/reciprocating behavior is the norm. Since in our stochastic model no strategy goes extinct, we have a very robust test of a norm's emergence. We hasten to add that, consistent with our interpretation of the results, our conclusions are also conditional. That is, they depend upon the strategy set we have considered. Given our strategy set, there is no unique evolutionarily stable strategy. Rather, we have a polymorphic population of strategies that uphold generalized trust and reciprocation.

REFERENCES

Axelrod, Robert (1984), *The Evolution of Cooperation.* New York: Basic Books.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995), "Trust, Reciprocity and Social History", *Games and Economic Behavior* 10(1): 122–142.

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

——— (forthcoming), *The Grammar of Society.* New York: Cambridge University Press.

Camerer, Colin (2003), *Behavioral Game Theory.* Princeton, NJ: Princeton University Press.

Cook, Karen (2001), *Trust in Society.* New York: Russell Sage Foundation.

Dasgupta, Partha (1988), "Trust as a Commodity", in Diego Gambetta (ed.), *Trust: Making and Breaking Cooperative Relations.* Oxford: Blackwell.

Fehr, Ernst, Simon Gachter, and Georg Kirchsteiger (1997), "Reciprocity as a Contract Enforcement Device: Experimental Evidence", *Econometrica* 65: 833–860.

Foster, Dean, and Peyton Young (1990), "Stochastic Evolutionary Game Dynamics", *Theoretical Population Biology* 38(2): 219–232.

Fudenberg, Drew, and Christopher Harris (1992), "Evolutionary Dynamics with Aggregate Shocks", *Journal of Economic Theory* 57(2): 420–441.

Fudenberg, Drew, and Eric Maskin (1986), "The Folk Theorem for Repeated Games with Discounting and Incomplete Information", *Econometrica* 54: 533–554.

Hofbauer, James, and Karl Sigmund (1998), *Evolutionary Games and Population Dynamics.* Cambridge: Cambridge University Press.

Kreps, David (1990), "Corporate Culture and Economic Theory", in James Alt and Karl Shepsle (eds.), *Perspectives on Positive Political Economy.* New York: Cambridge University Press, 90–143.

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

Maynard Smith, John, and Richard Price (1973), "The Logic of Animal Conflict", *Nature* 146: 15–18.

Putnam, Robert (1993), *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton, NJ: Princeton University Press.

——— (2000), *Bowling Alone: The Collapse and Revival of American Community.* New York: Simon and Schuster.

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

Samuelson, Larry, and Junfu Zhang (1992), ''Evolutionary Stability in Asymmetric Games'', *Journal of Economic Theory* 57(2): 363–391.

Skyrms, Brian (1990), *The Dynamics of Rational Deliberation.* Cambridge, MA: Harvard University Press.

——— (1996), *Evolution of the Social Contract.* New York: Cambridge University Press.

——— (1997), ''Chaos and the Explanatory Significance of Equilibrium: Strange Attractors in Evolutionary Game Dynamics'', in Cristina Bicchieri, Richard Jeffrey, and Brian Skyrms (eds.), *The Dynamics of Norms.* New York: Cambridge University Press.

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

Yamagishi, Toshio, and Midori Yamagishi (1994), ''Trust and Commitment in the United States and Japan'', *Motivation and Emotion* 18(2): 129–166.

Young, Peyton, and Dean Foster (1991), ''Cooperation in the Short and in the Long Run'', *Games and Economic Behavior* 3: 145–156.