*ISSN 1471-0498*

# DEPARTMENT OF ECONOMICS

# DISCUSSION PAPER SERIES

## LEARNING TO FORGIVE

**Thomas W.L. Norman**

Number 296

December 2006

Manor Road Building, Oxford OX1 3UQ

# Learning to Forgive*

Thomas W. L. Norman

All Souls College, Oxford OX1 4AL, UK

February 3, 2007

**Abstract**

The Folk Theorem for infinitely repeated games offers an embarrassment of riches; nowhere is equilibrium multiplicity more acute. This paper selects amongst these equilibria in the following sense. If players learn to play an infinitely repeated game using classical hypothesis testing, it is known that their strategies almost always approximate equilibria of the repeated game. It is shown here that if, in addition, they are sufficiently "conservative" in adopting their hypotheses, then almost all of the time is spent approximating an efficient subset of equilibria that share a "forgiving" property. This result provides theoretical justification for the general sense amongst practitioners that efficiency is focal in such games. *Journal of Economic Literature* Classification: C72; C12.

*Key Words:* Repeated games; Folk Theorem; learning; hypothesis testing; equilibrium selection.

## 1 Introduction

Game theorists' teeth are cut on the Prisoner's Dilemma:

| | Cooperate | Defect |
|---|---|---|
| Cooperate | 2<br>2 | 3<br>0 |
| Defect | 0<br>3 | 1<br>1 |

*I am grateful to Eric Maskin, David Myatt, Joe Perkins, Chris Wallace and Peyton Young for helpful discussions, and to seminar participants at Oxford University and Stony Brook 2006. Email thomas.norman@all-souls.ox.ac.uk.

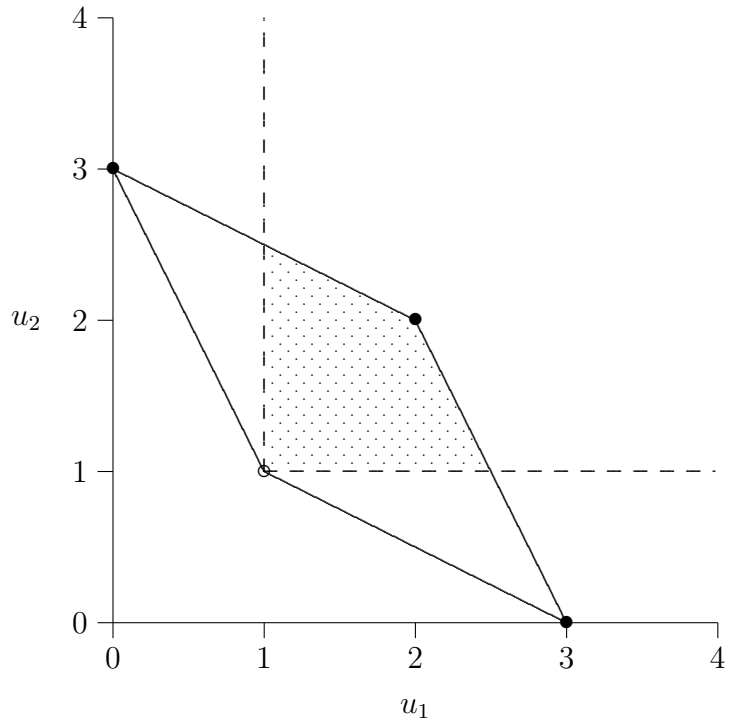


Figure 1: Equilibrium payoffs in the repeated Prisoner's Dilemma

With defection a dominant strategy, and thus the unique Nash equilibrium, we are left to wonder how players might cooperate, and thus realize a Pareto improvement. Repeating the game provides an intuitive and dramatic answer; the Folk Theorem for infinitely repeated games says that, if players are sufficiently patient, all feasible individually rational stage-game payoffs $(u_1, u_2)$ can be sustained in a (Nash or subgame-perfect) equilibrium of the repeated game (Aumann 1957, Aumann and Shapley 1976, Rubinstein 1979, Fudenberg and Maskin 1986). In Figure 1, the possibilities for the Prisoner's Dilemma are thus widened from the stage-Nash payoffs $(1, 1)$ to the entire shaded region, including efficient cooperation.

The Folk Theorem thus provides an answer to the puzzle of cooperation in the Prisoner's Dilemma, but it also clearly leaves us with a rather profound equilibrium-selection problem. Nevertheless, there is a general sense amongst practitioners that some of the equilibria attainable under the Folk Theorem are more appealing than others:

> "In applying repeated games, economists typically focus on one of the efficient equilibria, usually a symmetric one. This is due in part to a

> general belief that players may coordinate on efficient equilibria, and in part to the belief that cooperation is particularly likely in repeated games. It is a troubling fact that at this point there is no accepted theoretical justification for assuming efficiency in this setting."[1]

The idea of "renegotiation proofness" (Farrell and Maskin 1989, van Damme 1989, Pearce 1987, Abreu, Pearce, and Stacchetti 1993)—whereby a Pareto-dominated equilibrium in any subgame is "renegotiated" away—is one possible justification, but it sits a little uneasily with the noncooperative approach in general, and the criticisms of Pareto optimality in static games in particular.

An alternative justification for efficiency in repeated games is provided by the evolutionary approach. Axelrod's (1981, 1984) celebrated evolutionary simulations of the repeated Prisoner's Dilemma found selection pressure in favor of the strategy of "tit-for-tat," whereby a player cooperates in the first period and thereafter chooses the action his opponent took in the previous round. However, the outcome of such simulations is quite sensitive to the initial distribution of strategies upon which the selection process acts. On a theoretical level, meanwhile, the usual formulation of evolutionary stability suffers from severe existence problems in infinitely repeated games (Boyd and Lorberbaum 1987, Farrell and Ware 1988, Kim 1994), whilst a switch to neutral stability gives little sharpening of the predictions of the Folk Theorem. Nonetheless, Fudenberg and Maskin (1990, 1993) and Binmore and Samuelson (1992) do find efficiency to be implied by modified versions of neutral stability, and the arguments of the former will be important here.

An apparently attractive alternative avenue is provided by the literature on stochastic evolution (Foster and Young 1990, Kandori, Mailath, and Rob 1993, Young 1993, Ellison 2000), which offers up dynamic techniques for equilibrium selection that are insensitive to the initial distribution of strategies. The concept of stochastic stability picks out the equilibrium most likely to be played over the long-run evolution of a system made ergodic by the introduction of noise. However, such a system need not in general pick out a Nash equilibrium. Moreover, the learning interpretation of evolutionary models seems particularly strained in the case of repeated games; evolution requires a large number of repetitions of the *whole* game—a repeated repeated game, if you will—which is unappealing in many cases.

[1] Fudenberg and Tirole (1991), p. 160.

The learning literature seems to provide the more natural analytical framework of learning over the course of a *single* repeated game. Furthermore, whilst convergence to Nash equilibrium can be demanding in deterministic learning models (Kalai and Lehrer 1993, Nachbar 1997, Nachbar 2001, Nachbar 2005, Foster and Young 2001), Foster and Young (2003) have demonstrated probabilistic convergence to Nash (and indeed subgame perfection) in a stochastic model of learning by hypothesis testing. In this model, players hold hypotheses about their opponents' strategies, which they periodically test against observed play; if these are inconsistent, they reject their hypotheses in favor of alternatives. At any given time, players play a noisy best response to their current hypotheses, with some small probability that any action is played "by mistake." Aside from its empirical appeal, such a stochastic element is essential for convergence to Nash equilibrium in such "uncoupled" learning dynamics (Hart and Mas-Colell 2003, 2006).[2] Under these conditions, Foster and Young show that almost all time is spent approximating equilibria of the repeated game.

However, given the noise inherent in the hypothesis-testing process, it will not settle for so long on a *particular* equilibrium. Rather, any given equilibrium will be visited with a frequency determined by its attractiveness and persistence, as in the stochastic-stability literature. This paper investigates the implications of this observation for equilibrium selection in infinitely repeated games, under certain conditions on the hypothesis-testing process. In particular, it is found that, if two patient players are sufficiently "conservative" in revising their hypotheses—in the sense that a dramatic change in hypothesis is relatively unlikely—then the process spends most of its time approximating an efficient subset of equilibria of the repeated game. Conservatism is an appealing assumption in the present context, due both to the long-run nature of the analysis—where small modifications to beliefs should be expected—and to the fact that the hypothesis-testing strategies are $\epsilon$-best responses to beliefs in its presence. The selected equilibrium subset is defined by an attractive "forgiving" property shared by a common modification of the "tit-for-tat" strategy; finite numbers of mistakes do not lead to inefficiency.

The intuition for this equilibrium selection is that, with endogenous mistakes, equilibria where mistakes are not forgiven will eventually give inefficient payoffs. Moreover, there will eventually come a point where payoffs can be no worse in such an

[2]In an *uncoupled* dynamic, the adjustment of a player's strategy may not depend on the payoff functions of his opponents.

equilibrium. At this point, the players are free to experiment with efficient strategies, without the possibility of punishment; and they will thus find it optimal to do so even if there is only a small chance of reciprocation by their opponent. This feature—similar to the evolutionary-stability arguments of Fudenberg and Maskin (1990, 1993)—emerges naturally in the noisy setting of learning by hypothesis testing. If both players coincide in their experimentation, then efficient payoffs can result; and given long enough, such an escape from an "unforgiving" equilibrium becomes very likely. By contrast, in "forgiving" equilibria, payoffs remain efficient even following mistakes by the players, so that they are difficult to destabilize in this manner, requiring instead a dramatic change in hypotheses that is unlikely under conservatism. This combination of likely entry and unlikely exit means that the players will spend a lot of time approximating the set of forgiving equilibria in the long run. The main results of this paper make this intuition precise.

# 2 Informal Outline

The model of learning by hypothesis testing assumes that each player has a point estimate—a hypothesis—of the repeated-game strategies of all of the players. This in itself constitutes no departure from the standard framework of distributional beliefs; Kuhn's (1953) theorem tells us that any given probabilistic belief over behavior strategies can be replaced by some equivalent *particular* behavior strategy. Each player then chooses his own strategy as a "smoothed best response"—almost a best response, but with small probability placed on any strategy—to those of his opponents under his hypothesis. For example, in the infinitely repeated Prisoner's Dilemma, if a player hypothesizes that his opponent will always defect, then it is a smoothed best response for him to defect with high probability but cooperate with some small probability. It is thus possible for actions that do not constitute a best response to a player's hypothesis to be played "by mistake."

From time to time, each player will test his hypothesis against a sample of observed play, rejecting the hypothesis if it fails to fall within a certain tolerance level of the collected data. So if his hypothesis again has all players playing "always defect," a player might reject this hypothesis if the empirical distribution of the play that he observes places less than, say, 90% probability on this strategy profile.[3] In this

[3]Such a "naïve distance test" is, however, just one example of the broad class of hypothesis tests

case, he will adopt a new hypothesis and play a smoothed best response to that. Importantly, *any* new hypothesis is selected with some positive probability, and under this assumption Foster and Young (2003) demonstrate that almost all of the time is spent approximating equilibria of the repeated game.

One might, however, go further and argue that dramatic changes in hypotheses are unlikely from one period to the next, at least at the long-run level of analysis presented here. This is analogous to the evolutionary idea that an equilibrium should be robust to a small amount of mutation, with large mutations much less likely. This is captured in the hypothesis-testing model by the notion of "conservatism," under which a revised hypothesis lies "close" to the rejected hypothesis with high probability. The results in Section 4 will demonstrate that, under this additional assumption, almost all of the time is spent approximating not just equilibria, but an efficient class of equilibria that are "forgiving" in the sense that efficient play is preserved even following "mistakes" by the players.

The evolutionary stability of equilibria that are resilient to mistakes has been explored by Fudenberg and Maskin (1990). They incorporate infinitesimally probable mistakes explicitly into neutral stability, with players weighting events lexicographically in decreasing order of the number of mistakes required for the event to occur. When players employ finitely complex strategies and have time-average payoffs of this lexicographic form, Fudenberg and Maskin demonstrate that stability implies efficiency in the infinitely repeated Prisoner's Dilemma.[4] The essential idea is that, when players make mistakes, the worst possible initial history for an inefficient strategy profile will eventually occur; such a profile is then vulnerable to invasion by a mutant that mimics the incumbent strategy except after this worst history, at which point it engages in the familiar evolutionary "secret handshake" (Robson 1990), having nothing to lose from punishment. Mutants can then play efficiently amongst themselves, so that they prosper relative to incumbents. Similar possibilities arise naturally in the noisy setting of conservative hypothesis testing, as the following example illustrates.

Consider two players playing the infinitely repeated Prisoner's Dilemma according to (slightly noisy) "trigger" strategies: cooperate if and only if your opponent has

that might be carried out.

[4]In their 1993 working paper, the same authors extend this analysis to the discounting case with non-infinitesimal mistake probabilities.

always cooperated in the past. Suppose that one of the players has made a mistake, so that the players are locked in to perpetual defection. And suppose further that one of the players now tests and then rejects his current (correct) hypothesis in favor of a local alternative placing small probability on his opponent playing "perfect tit-for-tat"—whereby a player cooperates in the first period and thereafter cooperates if and only if either both players cooperated or both players defected in the previous period—starting in some particular future period $t'$. Then, if that player is sufficiently patient, it is a smoothed best response for him to "experiment" with perfect tit-for-tat from period $t'$: to also play perfect-tit-for-tat starting in $t'$, and continuing as long as there have been no more than $l$ deviations from perfect tit-for-tat since $t'$, where $l$ is a positive integer. To see this, note that it is optimal to play perfect tit-for-tat if it turns out that the opponent is playing perfect tit-for-tat, and if not, reversion to the trigger strategy after $l$ deviations yields only a small loss if the player is sufficiently patient. Crucially, things can be no worse for the player upon reversion to the trigger strategy, since play was already locked in to the worst possible scenario of perpetual defection.

Suppose now that the opponent too rejects his null hypothesis in favor of the same local alternative. Then it is a smoothed best response for him to experiment with perfect tit-for-tat starting in $t'$ in the same way. Cooperation may thus begin in period $t'$, and continue for some time in expectation if mistakes are unlikely. During this cooperative phase, if a player again rejects his hypothesis, and adopts a local alternative with somewhat more probability on his opponent playing perfect tit-for-tat, then it is a smoothed best response for that player to continue playing perfect tit-for-tat as long as there have been no more than $l'$ deviations from perfect tit-for-tat since $t'$, where $l' > l$; more mistakes can be forgiven, since it is now less likely that the opponent is playing the trigger strategy. Indeed, further model rejections and revisions may occur before reversion to trigger strategies takes place; and ultimately, enough model revisions may occur for the system to arrive in a state where both players are almost certain that their opponent is playing perfect tit-for-tat, a smoothed best response to which is to also play perfect tit-for-tat.

Once each player correctly believes that an equilibrium such as perfect tit-for-tat is being played, no local model revision can destabilize the associated efficiency. For there is no subgame where perfect tit-for-tat is strongly Pareto-dominated by some other equilibrium; it is an equilibrium that *forgives mistakes*. Alternative best re-

sponses in such equilibria also forgive mistakes in this manner, removing one source of instability. Moreover, whilst player $i$ can still place small probability on his opponent employing a response that would leave $i$ better off (e.g. the "always cooperate" strategy, against which defection goes unpunished), the opponent can have no incentive to adopt such a response, and $i$'s experimentation is thus doomed to break down, resulting in reversion to perfect tit-for-tat or an equivalent forgiving strategy. If such forgiving strategies are to be upset then, we require a (less probable) nonlocal model revision, or a sequence of (improbable) model rejections and revisions. Hence, the set of forgiving equilibria is relatively easy to enter and difficult to leave, so that the system will spend a lot of time there (or thereabouts).

This logic applies quite generally to favor equilibria that share the forgiving property of perfect tit-for-tat, as we will see in Section 4.

# 3 Hypothesis Testing

Now let us sketch the formal details of learning by hypothesis testing. There is a finite, $n$-person stage game $G$ with players $i = 1, 2, \ldots, n$, action spaces $X_i$ and utility functions $u_i : X \to \mathbb{R}$, $X = \Pi X_i$. The $u_i$'s are von Neumann–Morgenstern utility functions, normalized to lie between 0 and 1. This stage game is infinitely repeated in discrete time $t = 1, 2, \ldots$, with public observation of play. A *history of play* is denoted $\omega \in \Omega$; $\omega^t = (\omega_1^t, \ldots, \omega_n^t) \in X$ then denotes the actions taken in period $t$, $\overline{\omega}^t = (\omega^1, \omega^2, \ldots, \omega^t)$ the *initial history* of actions taken in periods 1 through $t$ inclusive, and $\Omega(\overline{\omega}^t) = \{\alpha \in \Omega \mid \overline{\alpha}^t = \overline{\omega}^t\}$ the set of all *continuations* of the initial history $\overline{\omega}^t$. Let $\overline{\Omega}(\overline{\omega}^t) = \{\overline{\alpha}^{t'} \mid t' \geq t,\ \overline{\alpha}^t = \overline{\omega}^t\}$ be the set of possible *continued initial histories* following $\overline{\omega}^{t-1}$.

Each player then has a forecast $p_i^t(x_{-i} \mid \overline{\omega}^{t-1}, b_i)$ of his opponents' one-step-ahead behaviors, conditional on every possible initial history, determined by his *model* $b_i$. Moreover, this model has *memory at most* $m$ in that the conditional distributions satisfy

$$p_i^t(x_{-i} \mid \overline{\omega}^{t-1}, b_i) = p_i^t(x_{-i} \mid \omega^{t-m}, \ldots, \omega^{t-1}, b_i) \quad \text{for all } t > m.$$

Since there are $M = |X|^m$ possible length-$m$ histories, models with memory at most $m$ occupy the Euclidean space $\mathcal{B}_i = \prod_{j \neq i} \Delta_j^M$. In response to his model, player $i$ adopts a *behavioral response* $a_i$ with *memory at most* $m$ in that the conditional probability

that $i$ plays action $x_i$ in period $t$, given the history $\overline{\omega}^{t-1}$, is[5]

$$q_i^t(x_i \mid \overline{\omega}^{t-1}, a_i) = q_i^t(x_i \mid \omega^{t-m}, \ldots, \omega^{t-1}, a_i) \quad \text{for all } t > m.$$

Note that $a_i \in \mathcal{A}_i = \Delta_i^M$, and $\mathcal{B}_i = \prod_{j \neq i} \mathcal{A}_j$. Letting $\mathcal{A} = \prod \mathcal{A}_i$ and $\vec{a} = (a_1, \ldots, a_n) \in \mathcal{A}$, we can define the mapping $B_i : \mathcal{A} \to \mathcal{B}_i$ from any response vector $\vec{a}$ to the correct model for $i$, $B_i(a) = \prod_{j \neq i} a_j$.

With a discount factor $\rho_i < 1$ for player $i$, $i$'s normalized discounted utility following the initial history $\overline{\omega}^{t-1}$ is $U_i^t(\omega) = (1 - \rho_i) \sum_{t'=t}^{\infty} \rho_i^{t'-t} u_i(\omega^{t'})$. Letting $\nu_{a_i, b_i}$ be the probability measure over infinite histories induced by the response $a_i$ and the model $b_i$, we can define $i$'s expected utility $U_i^t(a_i, b_i) \equiv \mathrm{E}\left(U_i^t(\omega) \mid a_i, b_i, \overline{\omega}^{t-1}\right)$ at time $t$ over all continuation histories $\Omega(\overline{\omega}^{t-1})$ as

$$\mathrm{E}\left(U_i^t(\omega) \mid a_i, b_i, \overline{\omega}^{t-1}\right) = \int_{\Omega(\overline{\omega}^{t-1})} U_i^t(\omega) d\nu_{a_i, b_i} \Big/ \int_{\Omega(\overline{\omega}^{t-1})} d\nu_{a_i, b_i}.$$

Given $\sigma_i > 0$, $a_i$ is then a *static $\sigma_i$-optimal response to $b_i$* if $U_i^t(a_i, b_i) \geq U_i^t(a_i', b_i) - \sigma_i$, $\forall a_i'$; if this holds for all $t$, then $a_i$ is an *extensive-form $\sigma_i$-optimal response to $b_i$*.[6] At any given time, each player $i$ has a (static or extensive-form) $\sigma_i$-optimal response function $A_i^{\sigma_i} : \mathcal{B}_i \to \mathcal{A}_i$ that is assumed to be *continuous* in $b_i$ and each payoff $u_i(x)$, and *diffuse* in the sense that each action is played with positive probability. Such an $A_i^{\sigma_i}$ is called a (static or extensive-form) *$\sigma_i$-smoothed best-response function* and $\{A_i^{\sigma_i} : \sigma_i > 0\}$ is a *family* of smoothed best-response functions. Let $\mathcal{S}^{\vec{\sigma}}$ be the set of fixed points of $A^{\vec{\sigma}} \circ B$, with $\mathcal{S}$ the set of (Nash or subgame-perfect, resp.) equilibrium response vectors obtained when $\sigma_i = 0$, $\forall i$.

Player $i$ periodically tests his *null hypothesis* that "the real process generating the actions from time $t$ on is described by the pair $(A_i^{\sigma_i}(b_i^t), b_i^t)$." If $i$ is not conducting a test at the start of a given period, he begins a new test with probability $1/s_i$. He then collects data on realized actions over the next $s_i$ periods, at the end of which he either accepts the null or rejects it. If it is rejected, he chooses a new model according to a probability measure $f_i^{t+s_i+1}(b_i \mid \overline{\omega}^{t+s_i})$. The hypothesis tests that the players employ are assumed to be *powerful*, in that the probability of player $i$ making a type-I or type-II error declines declines exponentially in $s_i$, given some small *tolerance level*

[5]For a discussion of such strategies, see Lehrer (1988).

[6]Foster and Young (2003) deal exclusively with extensive-form $\sigma_i$-optimality; for our purposes, the static formulation will represent an important and instructive foundation.

$\tau > 0$ for the test.

Clearly there will exist a great many alternative smoothed best responses to any given model. We will allow player $i$ to adopt any $\sigma_i$-smoothed best-response function $A_i^{\sigma_i}$ following the adoption of a new model. This constitutes a departure from the Foster and Young model, in which each player always employs the same $A_i^{\sigma_i}$. Once he has chosen such a response in our case though, he must still retain it until the next time he rejects his model.[7]

This is quite a general framework for learning. Indeed, the apparently quite different model of Bayesian learning in fact can be obtained as a limiting case of the hypothesis-testing model; where $s_i = 1$, $\tau \to 0$, $\sigma_i \to 0$ and $f_i^{t+s_i+1}(b_i \mid \overline{\omega}^{t+s_i})$ approaches a point mass on the $b_i'$ implied by Bayesian updating.

# 4 Conservatism and Forgiveness

The relevant formulation of the (perfect) Folk Theorem for this framework is Fudenberg and Maskin's (1986, 1991) discounting case, which also allows for unobservable mixed strategies.[8] The theorem says that, when players can observe each others' payoffs and actions, all feasible individually rational payoffs can be sustained in a subgame-perfect equilibrium of the infinitely repeated game if the discount factor is sufficiently close to one and a "full dimensionality" condition is satisfied. Foster and Young's (2003) results then say that, if players—lacking knowledge of their opponents' payoffs—engage in hypothesis testing about opponents' strategies, then there exists a range of parameters of this process such that the strategies are $\epsilon$-close to being a (Nash or subgame-perfect) equilibrium of the infinitely repeated game at least $1 - \epsilon$ of the time.

But we need not stop here; not all equilibria are equally appealing under hypothesis testing. Rather than settling on a particular equilibrium, for given $\epsilon > 0$ the process will perpetually bounce around between equilibria given long enough. But it will spend more time in some equilibria than in others, according to how likely they are to be entered and exited. The crucial variables in this respect are, for our purposes, the probabilities $f_i(\cdot|\overline{\omega}^t)$ with which new models are adopted upon hypothesis

[7] Modifying $\sigma_i$-optimality in this way is necessary in order to allow experimentation with efficiency.

[8] Also relevant will be Aumann and Shapley's (1976) time-average case, where players are arbitrarily patient.

rejection.

## 4.1 Conservatism

The basic assumption made on the hypothesis-revision densities $f_i(\cdot|\overline{\omega}^t)$ is *flexibility*, whereby, for each $\tau_0 > 0$, the $f_i$-measure of any $\tau_0$-ball of hypotheses is bounded below by a strictly positive number $f_*(\tau_0) > 0$. This is quite a weak assumption, allowing a wide range of possible models to be adopted at any given revision. In particular, absent further assumptions, it need not be more difficult for a player to adopt a model further away in the Euclidean model space $\mathcal{B}_i = \prod_{j \neq i} \Delta_j^M$, as would be the case in most evolutionary analyses. This role for the "distance" between models is instead captured by "conservatism," under which the new hypothesis lies within $\lambda_i$ of the old hypothesis with probability at least $1 - \lambda_i$, where $\lambda_i$ is positive and close to zero. Foster and Young do not use this assumption in their main convergence results; rather, they show that if players are sufficiently conservative, have sufficiently sharp best responses and employ sufficiently powerful hypothesis tests with sufficiently fine tolerances, then at all times the hypothesis-testing strategies are $\epsilon$-best responses to their *beliefs* (as distinct from models).[9]

We will modify conservatism slightly, in order to allow us to strengthen the concept somewhat.

**Definition 1** *Model revision is* conservative *if, following rejection, player $i$ adopts a new model that is within $\lambda_i$ of his previous model with probability at least $1 - \Lambda_i$.*

Conservatism thus still captures the idea that local model revisions are highly probable, but just how probable is no longer tied to the size of the neighborhood. By *sufficiently conservative*, we will mean that model revision is conservative with sufficiently small $\lambda_i$ and $\Lambda_i$.

## 4.2 Forgiveness

Once we allow individuals to choose any of the numerous alternative smoothed best responses available following hypothesis adoption, individual equilibria are doomed

[9]Beliefs differ from models in this setting because the players can anticipate and probabilize over their future model rejections. Beliefs are also unrestricted in their memory length.

to a common fragility. Hence, we consider the relative stability not of individual equilibria, but of a *set* of equilibria sharing a certain "forgiving" property.

A *state* $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is a hypothesis vector $(\vec{a}, \vec{b})$ following initial history $\overline{\omega}^{t-1}$. Given $\theta > 0$, let $\vec{\theta} = (\theta, \theta)$.

**Definition 2** *A state* $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ *is* weakly $\theta$-efficient *if* $(\mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1}))_{i=1,2}$ *is weakly Pareto-*$(\vec{\sigma} + \vec{\theta})$*-undominated in the set of equilibrium response vectors—i.e. if, for all equilibrium* $\vec{a}' \neq \vec{a}$*, there exists a player* $i$ *such that* $\mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1}) \geq \mathrm{E}(U_i^t(\omega)|a_i', B_i(\vec{a}'), \overline{\omega}^{t-1}) - (\sigma_i + \theta)$*. Otherwise, it is* strongly $\theta$-inefficient*.*

Note that we do not specify whether we mean Nash or subgame-perfect equilibrium here, since we will mean the former when dealing with static $\sigma_i$-optimality and the latter when extensive-form $\sigma_i$-optimality is introduced.

The *worst-case scenario* for player $i$ under $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is the initial history $\overline{\omega}_i^{t''-1} = \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}^{s_i}(\overline{\omega}^{t-1})} \mathrm{E}(U_i^{t'}(\omega)|a_i, b_i, \overline{\omega}^{t'-1})$.

**Definition 3** $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ *is* $\theta$-forgiving *if, for any worst-case scenario* $\overline{\omega}_i^{t''-1}$*,* $(\vec{a}, \vec{b}, \overline{\omega}_i^{t''-1})$ *is either weakly* $\theta$*-efficient or has probability at most* $\theta$ *following* $\overline{\omega}^{t-1}$*. Otherwise, it is* $\theta$-unforgiving*.*

Thus, if a state $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is $\theta$-forgiving, then any worst-case scenario such that some equilibrium response vector strongly Pareto-$(\vec{\sigma} + \vec{\theta})$-dominates $(U_i^{t'}(a_i, b_i))_{i=1,2}$ is reached with probability at most $\theta$. If $\theta = 0$, then we call $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ simply *forgiving* (and *weakly efficient*). Intuitively, $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is forgiving if any finite number of mistakes takes us no further than $\vec{\sigma}$ from efficiency. In the repeated Prisoner's Dilemma, perfect tit-for-tat following mutual cooperation (or defection) is forgiving in this sense. This is somewhat different (though related) to Axelrod's (1984, p. 36) informal notion of forgiveness in the Prisoner's Dilemma as the "propensity to cooperate in the moves after the other player has defected."

## 4.3 Equilibrium frequencies

Let us begin with a result for our static notion of $\sigma_i$-optimality, under which experimentation with forgiveness is easier to foster.

**Theorem 1** *Suppose that two players adopt hypotheses with finite memory, have* $\sigma_i$*-smoothed static best-response functions, employ powerful hypothesis tests, and are*

*flexible and conservative in the adoption of new hypotheses. Given any $\epsilon > 0$, if the $\sigma_i$ are small, if the test tolerance $\tau$ is sufficiently fine, if the amounts of data collected $s_i$ are sufficiently large, if the players are sufficiently conservative and patient, and if the memory length $m$ is sufficiently large, then the state of the process is $\epsilon$-forgiving at least $1 - \epsilon$ of the time.*

The proof of the theorem is relegated to the appendix.

The intuition behind the result is that, if a state is not forgiving, then it is vulnerable to experimentation initiated following the worst mistakes that will go unforgiven, since the experimenter has nothing to lose at this point and both players have something to gain. This is similar to the reasoning employed by Fudenberg and Maskin (1990), except that a one-off secret handshake followed by permanent efficiency or reversion can no longer work, since the continued possibility of mistakes makes the secret handshake's signal imprecise.[10] Instead, the players must first experiment a little, then gradually tolerate more and more mistakes as the experimentation is reciprocated, until they finally become completely efficient and forgiving.

On the other hand, if a state is forgiving, then experimentation cannot lead to a payoff profile strictly preferred by both players, and hence it will fail. Unlike in the symmetrized setting of Fudenberg and Maskin (1990), however, such failed experimentation can occur, and when it does the experimenter may revert not just to his previous response $a_i$, but to any alternative smoothed best response to his model. Indeed, either player may employ (an approximation of) an alternative smoothed best response upon hypothesis rejection even absent experimentation. Hence, any given response vector cannot be stable in the desired sense; rather, because a smoothed best response to a forgiving model must itself be $\theta'$-forgiving for some $\theta' \in (0, \epsilon)$, the set of $\epsilon$-forgiving response vectors is not exited for some long period of time. These observations are exploited to give the result on the long-run behavior of the dynamical system.

The idea of experimentation that is doomed to failure might seem a little odd; why should a player adopt a model (of reciprocated experimentation) under which his opponent fails to act in his own interests in such an elaborate manner? Should player $i$ not, for example, place low probability on his opponent playing a

[10] Fudenberg and Maskin (1993) do extend their analysis to non-infinitesimal mistake probabilities, but their arguments then require mistake probabilities to be very low—specifically, an order of magnitude lower than the discount rate.

response that is $\sigma_j$-optimal, $j \neq i$, under no possible models? The answer is that such reasoning can have no place in the hypothesis-testing model, since it would require players to have knowledge of opponents' payoffs and would thus sacrifice the "uncoupled" nature of the learning process.[11] In an uncoupled learning process, failed experimentation is necessarily a possibility. This underlines the probabilistic nature of the hypothesis-testing model; the delicate experimentation construction exploited to escape $\epsilon$-unforgiving states is not purposive or cooperative, but rather a coincidence of events that occur with positive probability. These events—and *a fortiori* their coincidence—are very unlikely, and hence take a long time to occur in expectation; but with sufficient conservatism, it takes even longer for the set of $\epsilon$-forgiving states to be exited, so that the result holds.

One deficiency of Theorem 1 is the static notion of $\sigma_i$-optimality employed. This facilitates experimentation, since it is not clear that Fudenberg and Maskin's worst-case scenario will endure long enough to sustain extended experimentation as $\sigma_i$-optimal in all subgames. Demonstrating that it will in fact do so if players are sufficiently patient is the key to the following result, which employs the extensive-form notion of $\sigma_i$-optimality.

**Theorem 2** *Suppose that two players now have $\sigma_i$-smoothed extensive-form best-response functions, but are otherwise as before. The conclusion of Theorem 1 again holds.*

The proof is again relegated to the appendix. The downside with this result is that it is likely to require a higher discount factor than Theorem 1; experimentation must now remain $\sigma_i$-optimal in every subgame where it continues, which requires a greater degree of patience on the part of the players.

With this extensive-form notion of $\sigma_i$-optimality, we are essentially back in the setting of Foster and Young (2003). The possibility of choosing alternative smoothed best responses in our model alters nothing important for their probabilistic convergence to equilibrium; disequilibrium states are still far more likely to be exited than equilibrium states given powerful hypothesis tests. Hence, combining their Theorem 2 with our Theorem 2 gives the following.

[11] On uncoupled learning processes, see Hart and Mas-Colell (2003, 2006), and Foster and Young (2006).

**Corollary 1** *Given $\epsilon > 0$, there exist values of the learning parameters such that the players' strategies are $\epsilon$-close to being an $\epsilon$-forgiving subgame-perfect equilibrium at least $1 - 2\epsilon$ of the time.*

# 5 Related Literature

The main power of these results comes when $\epsilon$ is small, in which case we can interpret them as selecting forgiving states from among the myriad possibilities of the Folk Theorem. The most directly comparable equilibrium-selection device in the literature is the notion of "stochastic stability" (Foster and Young 1990, Young 1993), which picks out those states receiving positive weight in a unique stationary distribution of play in the presence of vanishing noise. The techniques employed here, however, are quite distinct even from this evolutionary cousin. Most obviously, no limits are taken here, as we deal with (small but) positive levels of noise, indexed by the parameter $\epsilon > 0$. The distinction is stronger than this though; the stochastic process of learning by hypothesis testing is non-Markov—with transition probabilities able to depend on the entire history of play—and as a result we make no use of the standard ergodic apparatus for computing stochastically stable states. Instead, we work directly with the model's transition probabilities, bounding the likelihood of escaping forgiving and unforgiving states, and deriving the implications for their relative frequency under conservatism. In so doing, we exploit the link between conservatism—under which model revisions are local with high probability—and evolutionary stability, which essentially has local model revisions with probability one.

In its traditional formulation, however, evolutionary stability has had limited success in selecting between the equilibria possible under the various Folk Theorems. Axelrod and Hamilton (1981) show that "always defect" is not an ESS in the repeated Prisoner's Dilemma with time-average payoffs, since it is vulnerable to invasion by tit-for-tat (though this breaks down under discounting). Axelrod (1981, 1984) argues in favor of tit-for-tat on the basis of his concept of a "collectively stable strategy," but this concept does not imply evolutionary stability and gives little sharpening of the Nash Folk Theorem. Moreover, tit-for-tat is not a subgame-perfect equilibrium strategy against itself, and thus is not even a candidate equilibrium under the perfect Folk Theorems.

Boyd and Lorberbaum (1987) show that no pure strategy can be evolutionarily

stable in the infinitely repeated Prisoner's Dilemma, whilst Farrell and Ware (1988) extend this to finite mixtures of pure strategies. Kim (1994) generalizes these results to any strategies, and also to Selten's (1983) extensive-form concept of direct ESS. But Sugden (1986) and Boyd (1989) show that ESSs do exist if players occasionally make mistakes. The existence problem for direct ESS is the possibility of mutation to strategies that differ from the existing ones only off the equilibrium path. Selten's notion of a limit ESS addresses this problem by perturbing the game—so that every information set is reached with positive probability—and finding the limit of a sequence of direct ESSs as the perturbations vanish. This gives a refinement of sequential equilibrium in symmetric extensive-form games (van Damme 1987). However, Kim proves that a Folk Theorem obtains for limit ESSs; the concept offers no sharpening of the predictions of subgame perfection in the infinitely repeated Prisoner's Dilemma.

A similar criticism can be levelled at the relaxation of evolutionary stability to neutral stability, even with time-average payoffs, where there exist neutrally stable strategies of the infinitely repeated Prisoner's Dilemma that are arbitrarily close to "always defect" for example (Fudenberg and Maskin 1990). Modifications of evolutionary/neutral stability would thus seem to be required for significant refinements of the Folk Theorem. One such modification is Binmore and Samuelson's (1992) incorporation of complexity costs into neutral stability, which destabilizes the off-the-equilibrium-path punishments required to prevent secret handshakes, and thus provides selection pressure in favor of efficiency. Fudenberg and Maskin's (1990) notion of neutral stability with mistakes is another such modification, leading to selection pressure against strategies that harshly punish mistakes. In our model, similar pressures arise thanks to the close relation between neutral stability and conservatism; and given the endogenous noise in the hypothesis-testing process, we are able to exploit these pressures to provide probabilistic *convergence* results in a fully dynamic setting.

# 6 Conclusion

If two patient players learn to play an infinitely repeated game using classical hypothesis testing, and are sufficiently conservative in their adoption of new hypotheses, almost all time is spent approximating an efficient set of strategies that have an

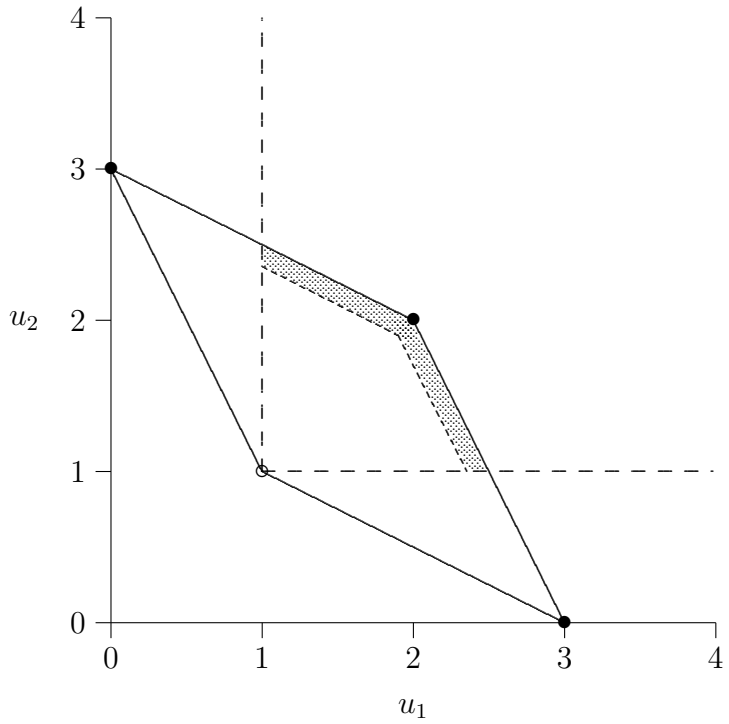


Figure 2: $\epsilon$-forgiving payoffs in the repeated Prisoner's Dilemma

intuitive forgiving property. For example, in the Prisoner's Dilemma, almost all time is spent close to the Pareto frontier of the feasible individually rational payoffs (the shaded region of Figure 2), enforced by equilibrium strategies such as perfect tit-for-tat that forgive finite numbers of mistakes. Intuitively, strategies that do not forgive mistakes are vulnerable to experimentation with efficiency once mistakes have been made. And whilst any *given* forgiving strategies are unstable in the face of alternative best replies, the *set* of forgiving strategies is stable; a change in long-run behavior requires both players to "agree" on their new behavior, which they cannot do if they are already playing efficiently.

# Appendix

**Time-average payoffs** A player's normalized *time-average payoff* from $\omega$ following $\overline{\omega}^{t-1}$ is $U_i^t(\omega) = \lim_{T\to\infty} \sum_{t'=t}^{T} u_i(\omega^{t'})/T$. Time-average payoffs are one possible formalization of the case where players do not discount the future, and they are insensitive to play in any finite number of periods (see, e.g., Osborne and Rubinstein 1994). If a history $\omega$ is preferred by player $i$ to another history $\omega'$ under time-average payoffs, then there is a discount factor $\rho_i$ close enough to 1 such that the same is true in the discounting case. These properties will be useful for our purposes.

The proofs of both theorems exploit the following lemma.[12]

**Lemma 1** *Under time-average payoffs, for any given hypothesis* $(a_i, b_i)$, $\mathrm{E}(U_i^{t'}(\omega)|a_i, b_i, \overline{\omega}^{t'-1})$ *is the same for all* $\overline{\omega}^{t'-1}$.

**Proof.** Suppose otherwise. Then there would exist a worst length-$m$ history and a strictly preferred continuation, a contradiction. ■

We call a $\theta$-forgiving state *generic* if a switch by either player to any given alternative best response under time-average payoffs gives another $\theta$-forgiving state. Whilst it is possible to construct $\theta$-forgiving states that are not generic in this sense, they are—as the name suggests—rather special. For an alternative best response for player $i$ in a $\theta$-forgiving state must give a new state with no subgame where player $i$'s expected (time-average) payoff is lower than in the old state; hence, to be $\theta$-unforgiving, this new state would have to have a subgame where player $i$'s expected payoff was the same as before, but that of player $j \neq i$ was lower.

**Proof of Theorem 1.** By way of preliminaries, let us recall some assumptions and resulting properties of the Foster and Young (2003) model that continue to hold in the present context. Choose all $\sigma_i$ such that

$$\forall i, \quad \sigma_i \leq \frac{\epsilon}{2}.$$

[12] I am grateful to Joe Perkins for pointing this out.

For any given $A_i^{\sigma_i}$,

$$\exists \delta > 0, \quad \forall \vec{u}, t, i, \quad \forall b_i, b_i', \\ |b_i - b_i'| \leq \delta \;\Rightarrow\; |A_i^{\sigma_i}(b_i) - A_i^{\sigma_i}(b_i')| \leq \frac{\epsilon}{4}, \quad \text{and} \quad \delta < \frac{\epsilon}{4}, \quad (1)$$

by continuity of $A_i^{\sigma_i}$. Fix $\tau \in (0, \delta)$; under powerful hypothesis tests, there exist functions $k(\tau)$ and $r(\tau)$ such that whenever a player's model is within $c(\tau)$ of the correct model, he rejects with probability at most

$$k(\tau)\mathrm{e}^{-r(\tau)s_*},$$

where $s_* := \min_i s_i$. In the proof of their Claim 2, Foster and Young demonstrate that there is a $\gamma > 0$ such that, for a given model fixed point $\vec{b}^f = B(\vec{a}^f)$,

$$\forall i, \quad \left| b_i - b_i^f \right| < \gamma \;\Rightarrow\; \left| b_i - B_i\left(A^{\vec{\sigma}}(\vec{b})\right) \right| < c(\tau).$$

Finally, recall the notion of a *great state*—where, for every player $i$, $i$'s model is within $\gamma$ of a fixed point and no player is currently in a test phase. If the fixed point is a forgiving state, we call any associated great state a *forgiving great state*.

Suppose to begin with that the process is in an $\epsilon$-unforgiving state $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ at time $t$. Then there exists a worst-case scenario $\overline{\omega}_1^{t''-1} = \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}^{s_1}(\overline{\omega}^{t-1})} \mathrm{E}(U_1^{t'}(\omega)|a_1, b_1, \overline{\omega}^{t'-1})$ such that $(\mathrm{E}(U_i^{t''}(\omega)|a_i, b_i, \overline{\omega}^{t''-1}))_{i=1,2}$ is strongly Pareto-$(\vec{\sigma} + \vec{\epsilon})$-dominated by some $\vec{a}^* \in \mathcal{S}$. Letting $\overline{\omega}_2^{t'''-1} := \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})} \mathrm{E}(U_2^{t'}(\omega)|a_2, b_2, \overline{\omega}^{t'-1})$, it follows that $\mathrm{E}(U_2^{t'''}(\omega)|a_2, b_2, \overline{\omega}^{t'''-1}) \leq \mathrm{E}(U_2^{t''}(\omega)|a_2, b_2, \overline{\omega}^{t''-1})$. Let the response $a_1^{l_1}$ continue to play $a_1$ unless $\overline{\omega}_2^{t'''-1}$ is realized, in which case it plays $a_1^*$ if and only if there have been at most $l_1$ deviations from $\vec{a}^*$ (by either player) since $t'''$; otherwise, it reverts to $a_1$ forever. This response clearly does not have memory $m$—and in fact requires infinite memory over the whole game—but we can always choose $m$ sufficiently large such that if $a_1^{l_1}$ is $\sigma_1'$-optimal, $\sigma_1' < \sigma_1$, given some memory-$m$ model, then there exists a $\sigma_1$-optimal memory-$m$ response constructed from $a_1^{l_1}$ using the procedure described in Foster and Young (2003, p. 81). Similarly, let $a_2^{l_2}$ respond to $\overline{\omega}_2^{t'''-1}$ by playing $a_2^*$ if and only if there have been at most $l_2$ deviations from $\vec{a}^*$ since $t'''$, otherwise reverting to $a_2$ forever. Finally, let $a_i^\infty := \lim_{l_i \to \infty} a_i^{l_i}$ be the response that plays $a_i^*$ for all $\overline{\omega}^{t'-1}$ once adopted, and $b_i' = (1 - \lambda_i + \iota_i)b_i + (\lambda_i - \iota_i)a_j^\infty$, $\iota_i < \min\{\delta, \lambda_i/2\}$, $j \neq i$. Consider the following

sequence of events leading to a generic forgiving great state with fixed point $\vec{a}^{\infty}$.

**Step 1.** Play proceeds in accordance with $\overline{\omega}_1^{t''-1}$. Player 2 does not start a test between periods $(t'' - \max\{s_1, 2s_2\})$ and $(t'' - 1)$; player 1 does not start a test between periods $(t'' - 2s_1)$ and $(t'' - (s_1 + 1))$, but does so in period $(t'' - s_1)$. After 1's test phase is completed, he rejects his current hypothesis and adopts a model within $\iota_1$ of $b_1'$. Now, fixing $\sigma_1' \in ((1 - \lambda_1 + \iota_1)\sigma_1, \sigma_1)$, whilst $\mathrm{E}(U_1^{t'}(\omega)|a_1^{l_1}, b_1, \overline{\omega}_1^{t'-1})$ is nonincreasing in $l_1$ for given $\rho_1 < 1$ and any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$, if $\rho_1$ is sufficiently high there exists a maximum $\bar{l}_1 > 0$ such that $a_1^{\bar{l}_1}$ is a $\sigma_1^+$-optimal response to $b_1$ for all $l_1 \leq \bar{l}_1$ and $\sigma_1^+ := \sigma_1'/(1 - \lambda_1 + \iota_1) > \sigma_1$. To see this, note that, by eventually reverting to $a_1$, $a_1^{l_1}$ gives the same time-average payoff as $a_1$ against $b_1$ following $\overline{\omega}_1^{t''-1}$. And since $a_1$ is a $\sigma_1$-optimal response to $b_1$, it follows that $a_1^{l_1}$ must be a $\sigma_1^+$-optimal response to $b_1$ under discounting given $l_1$ sufficiently low and $\rho_1$ sufficiently high. Moreover, since $a_1^*$ is an optimal response to $a_2^*$, mistakes have zero probability under $\vec{a}^*$ and hence $a_1^{\bar{l}_1}$ must be an optimal response to $a_2^{\infty}$. Hence,

$$\begin{aligned}
\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, b_1', \overline{\omega}_1^{t''-1}) &= (1 - \lambda_1 + \iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, b_1, \overline{\omega}_1^{t''-1}) \\
&\qquad + (\lambda_1 - \iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, a_2^{\infty}, \overline{\omega}_1^{t''-1}), \\
&\geq (1 - \lambda_1 + \iota_1)(\sup_{a_1'} \mathrm{E}(U_1^{t''}(\omega)|a_1', b_1, \overline{\omega}_1^{t''-1}) - \sigma_1^+) \\
&\qquad + (\lambda_1 - \iota_1) \sup_{a_1'} \mathrm{E}(U_1^{t''}(\omega)|a_1', a_2^{\infty}, \overline{\omega}_1^{t''-1}), \\
&\geq \sup_{a_1'} \mathrm{E}(U_1^{t''}(\omega)|a_1', b_1', \overline{\omega}_1^{t''-1}) - (1 - \lambda_1 + \iota_1)\sigma_1^+ \\
&= \sup_{a_1'} \mathrm{E}(U_1^{t''}(\omega)|a_1', b_1', \overline{\omega}_1^{t''-1}) - \sigma_1',
\end{aligned}$$

so that $a_1^{\bar{l}_1}$ is a $\sigma_1'$-optimal response to $b_1'$. (1) then implies that player 1 has a $\sigma_1$-optimal response to his new model within $\epsilon/4$ of (a $\sigma_1$-optimal memory-$m$ response to $b_1'$ appropriately constructed from) $a_1^{\bar{l}_1}$. (Duration: $(t'' - t)$ periods.)

**Step 2.** Play proceeds in accordance with $\overline{\omega}_2^{t'''-1}$. In period $(t''' - s_2)$, player 2 starts a test period, at the end of which he rejects and adopts a model within $\iota_2 < \delta, \lambda_2/2$ of $b_2'$. Since $\mathrm{E}(U_2^{t'''}(\omega)|a_2, b_2, \overline{\omega}^{t'''-1}) \leq \mathrm{E}(U_2^{t''}(\omega)|a_2, b_2, \overline{\omega}^{t''-1})$, the argument in Step 1 then applies, so that there is a $\sigma_2$-optimal response to $b_2'$ within $\epsilon/4$ of (a $\sigma_2$-optimal

memory-$m$ response to $b_2'$ appropriately constructed from) $a_2^{\bar{l}_2}$. (Duration: $(t''' - t'')$ periods.)

**Step 3.** Each player $i$ conducts successive non-overlapping tests, rejecting at the end of each and adopting a $\lambda_i$-close model within $\iota_i$ of a linear combination of $b_i$ and $a_j^\infty$, $j \neq i$, that maximizes the weight on $a_j^\infty$ subject to its $\iota_i$-ball being contained in the rejected model's $\lambda_i$-ball, until he adopts a model within $\min\{\gamma, \delta\}$ of $a_j^\infty$. Throughout this process, the players adopt approximations of $a_i^{\bar{l}_i}$ with successively higher $\bar{l}_i$'s (which are $\sigma_i$-optimal by the arguments in Step 1), and there are sufficiently few deviations from $\vec{a}^*$ to provoke reversion to $\vec{a}$. Once the players' hypotheses approximate $\vec{a}^\infty$, (1) implies the existence of a $\sigma_i$-optimal memory-$m$ response within $\epsilon/4$ of $a_i^\infty$. (Duration: at most $\varrho := \lceil 2(1 - \min\{\gamma, \delta\} - \lambda_* + 2\iota^*)/(\lambda_* - 2\iota^*)\rceil s^*$ periods, where $\lambda_* := \min_i \lambda_i$, $\iota^* := \max_i \iota_i$ and $s^* := \max_i s_i$.)

**Step 4.** If the number of periods in Steps 1–3 is $T' < t''' - t + \varrho$, no player begins a test for the next $t''' - t + \varrho - T'$ periods.

The duration of the whole sequence is exactly $t''' - t + \varrho$.

We now calculate the probability of a particular such sequence. To begin with, $\overline{\omega}_1^{t''-1}$ must be realized, which occurs with probability $\epsilon$ or more. There must be no rejections between $t$ and $t''-1$, which occurs with probability at least $(1-1/s_*)^{2(t''-s_*-t)}$, where $s_* := \min_i s_i$. Player 1's first test phase must then begin in period $(t'' - s_1)$, and player 2 must not test during this phase, which occurs with probability at least $(1/s^*)(1 - 1/s_*)^{s^*}$. Player 1 must then reject his null hypothesis, and adopt a new one within a target of radius $\iota_1$ in his model space, which occurs with probability at least $(1 - \nu^*)f_*$, where $f_* = f_*(\iota_*)$ and $\nu^*$ is the maximum probability of a player accepting his null hypothesis. He must then adopt the relevant smoothed best-response function from amongst the set of alternatives, which occurs with probability at least $\xi_*$ say.

Next, $\overline{\omega}_2^{t'''-1}$ must be realized, which occurs with some positive probability, $\phi$ say. There must be no rejections between $t''$ and $t''' - 1$, followed by a player-2 test phase starting at $(t''' - s_2)$ and with no simultaneous testing by player 1, which occurs with probability at least $(1/s^*)(1 - 1/s_*)^{2(t'''-1-s_*-t'')+s^*}$. Player 2 must then reject his null hypothesis, adopt a new one within a target of radius $\iota_2$ in his model space,

and select the relevant smoothed best-response function; this event has probability at least $(1-\nu^*)f_*\xi_*$.

The Step-3 test phases must each begin at a specific time and no player can be testing during the other's phase; the probability of this is at least $\left((1/s^*)(1-1/s_*)^{s^*}\right)^{\varrho/s^*}$. Each of these tests must end with rejection, subsequent adoption within a model-space target of radius $\iota_i$, and selection of the relevant smoothed best-response function; we can choose the test parameters such that the rejections occur with probability at least $1/2$ (since the players' models now place relatively low probability on the realized path of play), so that the event has probability at least $(f_*\xi_*/2)^{\varrho/s^*}$. There must be sufficiently few deviations from $\vec{a}^*$ from $\overline{\omega}_2^{t'''-1}$ to the end of Step 4 to avoid reversion to $\vec{a}$; let the probability of this event be $\psi$.

Finally, there must be no further tests before period $t''' + \varrho$; this occurs with probability at least $(1-1/s_*)^{2(t''-t+\varrho)}$.

In summary, the probability of Steps 1–4 is at least

$$\epsilon\phi\psi(1-\nu^*)^2(1-1/s_*)^{3\varrho+2(t'''+t''-1-2t+s^*-2s_*)}(f_*\xi_*/2s^*)^{2+\varrho/s^*}.$$

Thus there are constants $\alpha, \beta \in (0,1)$ such that the probability of Steps 1–4 is at least $\alpha\beta^\varrho$, establishing the following lemma.

**Lemma 2** *If $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is $\epsilon$-unforgiving, then there exist constants $\alpha, \beta \in (0,1)$, $\rho_i$ sufficiently close to 1 and $m$ sufficiently high such that the probability of being in a generic forgiving great state at time $t''' + \varrho$ is at least $\alpha\beta^\varrho$.*

Lemma 2 bounds the likelihood of entering a generic forgiving great state, but to prove Theorem 1 we need also to say something about exiting such states. Let $\Lambda^* := \max_i \Lambda_i$, $t^{\max}$ be the maximum over all states of the time interval required to effect both players' worst remaining initial histories in the most probable manner, and $\lceil x \rceil := \min\{z \in \mathbb{Z} \mid z \geq x\}$.[13]

**Lemma 3** *If $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is a generic forgiving great state, then there exists a constant $\eta \in (0,1)$, $\rho_i$'s sufficiently close to 1, $\lambda_i$'s sufficiently low and $s_i$'s sufficiently high such that the probability of being in an $\epsilon$-unforgiving state within $T$ periods is at most $\eta\Lambda^* T$, where $T = \lceil (t^{max}+\varrho)(1+\beta^{-2\varrho})/\epsilon \rceil$ and $\Lambda^* = \epsilon\tilde{\beta}^{3\varrho}/(t^{max}+\varrho)$ for some constant $\tilde{\beta} \in (0,1)$.*

[13] This is finite, given models and responses of memory at most $m$ and the finitely many possible length-$m$ histories.

**Proof.** Since the process begins in a great state at time $t$, each null hypothesis $(a_i^t, b_i^t)$ is within $\tau_0 = c(\tau)$ of the truth. And since $T$ is a large positive integer, the probability that any player rejects a test over the next $T$ periods is bounded above by

$$\lceil 2T/s_* \rceil k_0 \mathrm{e}^{-r_0 s_*} < T\mathrm{e}^{-4rs_*},$$

where $k_0 = k(\tau_0) > 0$, $r_0 = r(\tau_0) > 0$ and the inequality holds for all sufficiently large $s_*$ and some $r > 0$. Because a dramatic change in hypotheses then occurs with probability at most $\Lambda^*$, it follows that there is a constant $\eta \in (0,1)$ such that the probability of being in an $\epsilon$-unforgiving state within $T$ periods is at most $\eta\Lambda^* T$. To establish the result, we must show that escaping the $\epsilon$-forgiving states through local model revisions is less likely than this.

We start by claiming that, since $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is a generic forgiving state, there exists $\theta' > 0$ and sufficiently small $\lambda_i$'s such that, $\forall \theta \geq \theta'$, no player $i$ has a model $b_i' = (1-\zeta)b_i + \zeta b_i''$, $\zeta \leq \lambda_i$, $b_i'' \in \mathcal{B}_i$, that can induce a $\sigma_i$-optimal response $a_i'$ such that the resulting state is $\theta$-unforgiving. To see this note that, for any $\theta > 0$, if a state is $\theta$-unforgiving under time-average payoffs, then there exists a subgame where it is strongly $\theta$-inefficient; hence it is strongly $\theta$-inefficient in *all* subgames by Lemma 1, and in particular it is strongly $\theta$-inefficient in the *current* subgame. Hence, for all $\theta > 0$, there exists an equilibrium response vector that strongly Pareto-$\vec{\sigma}$-dominates it in the current subgame. Choose the $\lambda_i$'s sufficiently small that $\vec{a}$ is such a vector. This is possible because $\mathrm{E}(U_i^t(\omega)|a_i', b_i', \overline{\omega}^{t-1}) \to \mathrm{E}(U_i^t(\omega)|a_i', b_i, \overline{\omega}^{t-1})$ as $\lambda_i \to 0$, and $\mathrm{E}(U_i^t(\omega)|a_i', b_i, \overline{\omega}^{t-1}) < \mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1})$ since $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is generic; $(\vec{a}, \vec{b}, \overline{\omega}^{t-1})$ is forgiving and hence weakly efficient (i.e. $(\mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1}))_{i=1,2}$ is weakly Pareto-$\vec{\sigma}$-undominated in the set of equilibrium response vectors). But then $\mathrm{E}(U_i^t(\omega)|a_i', b_i', \overline{\omega}^{t-1}) < \mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1}) - \sigma_i$ and, for sufficiently small $\lambda_i$, $\mathrm{E}(U_i^t(\omega)|a_i', b_i, \overline{\omega}^{t-1}) < \mathrm{E}(U_i^t(\omega)|a_i, b_i, \overline{\omega}^{t-1}) - \sigma_i$; $a_i'$ cannot be a $\sigma_i$-optimal response to $b_i$. It follows that there exists some $\theta' > 0$ such that the same will be true under discounting for given $\rho_i < 1$ and all $\theta \geq \theta'$.

The same argument implies that, for a generic $\theta'$-forgiving state, there exists $\theta'' > \theta'$ and sufficiently small $\lambda_i$'s such that, $\forall \theta \geq \theta''$, no player $i$ has a model $b_i' = (1-\zeta)b_i + \zeta b_i''$, $\zeta \leq \lambda_i$, $b_i'' \in \mathcal{B}_i$, that can induce a $\sigma_i$-optimal response $a_i'$ such that the resulting state is $\theta$-unforgiving. Since $\theta'$ and $\theta'' - \theta'$ vanish as $\rho_i \to 1$, for given $\lambda_i$'s we can then choose the $\rho_i$'s sufficiently close to 1 that escaping the set

of $\epsilon$-forgiving states is more likely through a dramatic change in hypotheses than through local revisions. Intuitively, this is because having the $\rho_i$'s sufficiently close to 1 necessitates multiple unlikely hypothesis rejections by each player in order to reach an $\epsilon$-unforgiving state. The rejections are unlikely because discontinuities in $\sigma_i$-optimal behavior cannot endure for long in a forgiving state; experimentation is doomed to failure and reversion to former behavior or a close approximation. And if behavior remains close to former behavior, further hypothesis rejections remain unlikely. In particular, alternative best responses must differ in at most finitely many periods under time-average payoffs; hence, by choosing the $\rho_i$'s and $s_*$ sufficiently high, we can ensure that future samples of play provoking hypothesis rejection are sufficiently unlikely. ■

We can now use the above bounds to show that the fraction of times that the process is not in an $\epsilon$-forgiving state is very small if players are sufficiently patient and conservative. Starting from time $t$, let $\mathcal{E}$ be the event "the realized states in at least $\epsilon T$ of the periods $t+1,\ldots,t+T$ are $\epsilon$-unforgiving." Let $\mathcal{E}'$ be the sub-event of $\mathcal{E}$ in which no generic forgiving great state is realized before the last $\epsilon$-unforgiving state, and let $\mathcal{E}'' = \mathcal{E} - \mathcal{E}'$. We shall bound the conditional probabilities of $\mathcal{E}'$ and $\mathcal{E}''$ from above independently of the state at time $t$.

Fix $s_i$'s sufficiently high, $\lambda_i$'s sufficiently low, $\rho_i$'s sufficiently close to 1 and $m$ sufficiently high that Lemmas 2 and 3 both hold with $\tilde{\beta} = \beta$, $\beta^{-\varrho} > -(1/\alpha)\ln(\epsilon/2)$ and $\beta^{3\varrho} + \beta^{\varrho} < \epsilon/2\eta$. Since Lemma 2 is true for any given $s_i$'s, they can be chosen sufficiently high for Lemma 3 to hold; since Lemma 3 holds for all sufficiently low $\lambda_i$'s, we can choose them such that $\beta^{-\varrho} > -(1/\alpha)\ln(\epsilon/2)$ and $\beta^{3\varrho} + \beta^{\varrho} < \epsilon/2\eta$, given the $s_i$'s; since Lemma 2 is true for all $\rho_i$'s sufficiently close to 1, they can be chosen high enough to give consistency with Lemma 3, given the $s_i$'s and $\lambda_i$'s; and $m$ can then be chosen sufficiently high for the approximations of Lemma 2 to hold.

If $\mathcal{E}'$ occurs, there are at least $\lfloor \epsilon T/(t^{\max} + \varrho) \rfloor = k$ distinct times $t < t_1 < \cdots < t_k \leq t + T$ such that the following hold:

- $t_{j+1} - t_j \geq t^{\max} + \varrho$ for $1 \leq j < k$,
- the state at time $t_j$ is $\epsilon$-unforgiving for $1 \leq j < k$,
- no generic forgiving great state occurs from $t_1$ to $t_k$.

By Lemma 2, the probability of this event is at most $(1-\alpha\beta^{\varrho})^{k-1} \leq \mathrm{e}^{-\alpha\beta^{\varrho}(k-1)}$. Since

$T = \lceil (t^{\max} + \varrho)(1 + \beta^{-2\varrho})/\epsilon \rceil$, we have

$$P(\mathcal{E}') \leq \exp(-\alpha\beta^{\varrho}(\lfloor \epsilon T/(t^{\max} + \varrho) \rfloor - 1)) = \exp(-\alpha\beta^{-\varrho}),$$

where $\lfloor x \rfloor := \max\{z \in \mathbb{Z} \mid z \leq x\}$. Since $\beta^{-\varrho} > -(1/\alpha)\ln(\epsilon/2)$, it follows that $P(\mathcal{E}')$ is less than $\epsilon/2$.

If $\mathcal{E}''$ occurs, the process does *not* stay in $\epsilon$-forgiving states for at least $T$ periods after entering a generic forgiving great state. So from Lemma 3, and since $\Lambda^* = \epsilon\beta^{3\varrho}/(t^{\max} + \varrho)$,

$$P(\mathcal{E}'') \leq \eta\Lambda^* T = \eta\Lambda^*(t^{\max} + \varrho)(1 + \beta^{-2\varrho})/\epsilon = \eta(\beta^{3\varrho} + \beta^{\varrho}),$$

which is also less than $\epsilon/2$ since $\beta^{3\varrho} + \beta^{\varrho} < \epsilon/2\eta$. Putting all of this together we conclude that, for all sufficiently small $\lambda_*$,

$$P(\mathcal{E}) = P(\mathcal{E}') + P(\mathcal{E}'') \leq \epsilon.$$

Now divide all times $t$ into disjoint blocks of length $T$, and let $Z_k$ be the fraction of $\epsilon$-unforgiving times in the $k$th block. We have just shown that $P(Z_k \geq \epsilon) \leq \epsilon$ for all $k$. Hence

$$E(Z_k) \leq P(Z_k \geq \epsilon) \cdot 1 + P(Z_k < \epsilon) \cdot \epsilon \leq 2\epsilon.$$

It follows that the proportion of times that the process is in an $\epsilon$-unforgiving state is almost surely less than $2\epsilon$. Rerunning the entire argument with $\epsilon/2$ yields the desired conclusion, namely that the state of the process is $\epsilon$-forgiving at least $1 - \epsilon$ of the time. ■

**Proof of Theorem 2.** Let $\tilde{a}_i^{l_i}$ be the modification of $a_i^{l_i}$ that, in the event of more than $l_i$ deviations from $\vec{a}^*$, switches to an optimal response to $b_1$—$\tilde{a}_1$ say—rather than reverting to $a_1$. The proof proceeds in the same way as that of Theorem 1, except that in Step 1 on p. 20, $b_i'$ is replaced by $b_i^l = (1 - \lambda_i + \iota_i)b_i + (\lambda_i - \iota_i)a_j^l$, $\iota_i < \min\{\delta, \lambda_i/2\}$, $j \neq i$, for some $l \in \mathbb{N}$ such that $\tilde{a}_i^l$ is an extensive-form $\sigma_i'$-optimal response to $b_i^l$, $\sigma_i' < \sigma_i$.[14] To see that such an $l$ exists, note that $\tilde{a}_1^l$ is an optimal response to $a_2^l$,

[14] We now have the added complication that $b_i^l$ does not have memory $m$, but since an appropriate memory-$m$ approximation approaches $b_i^l$ as $m$ becomes large, the results again hold for $m$ sufficiently large.

and fix $\sigma_1' \in ((1-\lambda_1+\iota_1)\sigma_1, \sigma_1)$. Whilst $\mathrm{E}(U_1^{t'}(\omega)|\tilde{a}_1^{l_1}, b_1, \overline{\omega}_1^{t'-1})$ is nonincreasing in $l_1$ for given $\rho_1 < 1$ and any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$, if $\rho_1$ is sufficiently high there exists a maximum $\bar{l}_1 > 0$ such that $\tilde{a}_1^{\bar{l}_1}$ is a $\sigma_1^+$-optimal response to $b_1$ for all $l_1 \leq \bar{l}_1$ and $\sigma_1^+ := \sigma_1'/(1-\lambda_1+\iota_1) > \sigma_1$. To see this recall that, under time-average payoffs, for any given hypothesis $(a_i, b_i)$, $\mathrm{E}(U_i^{t'}(\omega)|a_i, b_i, \overline{\omega}^{t'-1})$ is the same for all $\overline{\omega}^{t'-1}$ by Lemma 1. Hence, by eventually switching to $\tilde{a}_1$, $\tilde{a}_1^{l_1}$ gives at least the same time-average payoff as $a_1$ against $b_1$ following $\overline{\omega}^{t''-1}$, and indeed following any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$. And since $a_1$ is a $\sigma_1$-optimal response to $b_1$, it follows that $\tilde{a}_1^{l_1}$ must be a $\sigma_1^+$-optimal response to $b_1$ under discounting given $l_1$ sufficiently low and $\rho_1$ sufficiently high. Hence, for $l \leq \bar{l}_1$,

$$\begin{aligned}
\mathrm{E}(U_1^{t''}(\omega)|\tilde{a}_1^l, b_1^l, \overline{\omega}_1^{t''-1}) &= (1-\lambda_1+\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|\tilde{a}_1^l, b_1, \overline{\omega}_1^{t''-1}) \\
&\qquad + (\lambda_1-\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^l, a_2^l, \overline{\omega}_1^{t''-1}), \\
&\geq (1-\lambda_1+\iota_1)(\sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1, \overline{\omega}_1^{t''-1}) - \sigma_1^+) \\
&\qquad + (\lambda_1-\iota_1)\sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, a_2^l, \overline{\omega}_1^{t''-1}), \\
&\geq \sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1^l, \overline{\omega}_1^{t''-1}) - (1-\lambda_1+\iota_1)\sigma_1^+ \\
&= \sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1^l, \overline{\omega}_1^{t''-1}) - \sigma_1',
\end{aligned}$$

so that $\tilde{a}_1^l$ is a $\sigma_1'$-optimal response to $b_1^l$. A similar sequence of events to that in Theorem 1 then leads to a great state near the fixed point $\vec{a}^l$, which is $\theta'$-forgiving for some $\theta' > 0$. Increasing the $\rho_i$'s increases the value of $l$ consistent with $\sigma_i$-optimality, and thus reduces $\theta'$; hence, we can choose the $\rho_i$'s sufficiently close to 1 that the result holds. ■

# References


Abreu, D., D. Pearce, and E. Stacchetti (1993): "Renegotiation and Symmetry in Repeated Games," *Journal of Economic Theory*, 60, 217–240.

Aumann, R. J. (1957): "Acceptable Points in General Cooperative $n$-Person Games," in *Contributions to the Theory of Games IV,* Annals of Mathematics Study 40, ed. by R. D. Luce, and A. W. Tucker, pp. 287–324. Princeton University Press, Princeton NJ.

Aumann, R. J., and L. Shapley (1976): "Long-Term Competition—A Game-Theoretic Analysis," Mimeo, Hebrew University. Reprinted in *Essays in Game Theory*, ed. by N. Megiddo (1994). Springer-Verlag, New York.

Axelrod, R. (1981): "The Emergence of Cooperation Among Egoists," *American Political Science Review*, 75, 306–318.

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

Axelrod, R., and W. Hamilton (1981): "The Evolution of Cooperation," *Science*, 211, 1390–1396.

Binmore, K. G., and L. Samuelson (1992): "Evolutionary Stability in Repeated Games Played by Finite Automata," *Journal of Economic Theory*, 57, 278–305.

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

Boyd, R., and J. Lorberbaum (1987): "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoners' Dilemma Game," *Nature*, 327, 58–59.

Ellison, G. (2000): "Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution," *Review of Economic Studies*, 67, 17–45.

Farrell, J., and E. Maskin (1989): "Renegotiation in Repeated Games," *Games and Economic Behavior*, 1, 327–360.

Farrell, J., and R. Ware (1988): "Evolutionary Stability in the Repeated Prisoner's Dilemma Game," *Theoretical Population Biology*, 36, 161–166.

Foster, D. P., and H. P. Young (1990): "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, 38, 219–232.

——— (2001): "On the Impossibility of Predicting the Behavior of Rational Agents," *Proceedings of the National Academy of Sciences of the USA*, 98(22), 12848–12853.

——— (2003): "Learning, Hypothesis Testing, and Nash Equilibrium," *Games and Economic Behavior*, 45, 73–96.

——— (2006): "Regret Testing: Learning to Play Nash Equilibrium without Knowing You Have an Opponent," *Theoretical Economics*, 1, 341–367.

Fudenberg, D., and E. Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, 54, 533–554.

——— (1990): "Evolution and Cooperation in Repeated Games," *American Economic Review Papers and Proceedings*, 80, 274–279.

——— (1991): "On the Dispensability of Public Randomization in Discounted Repeated Games," *Journal of Economic Theory*, 53, 428–438.

——— (1993): "Evolution and Repeated Games," Working Paper.

Fudenberg, D., and J. Tirole (1991): *Game Theory*. The MIT Press, Cambridge, Massachusetts.

Hart, S., and A. Mas-Colell (2003): "Uncoupled Dynamics Do Not Lead to Nash Equilibrium," *American Economic Review*, 93, 1830–1836.

——— (2006): "Stochastic Uncoupled Dynamics and Nash Equilibrium," *Games and Economic Behavior*, 57, 286–303.

Kalai, E., and E. Lehrer (1993): "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61, 1019–1045.

Kandori, M., G. J. Mailath, and R. Rob (1993): "Learning, Mutation and Long-Run Equilibria in Games," *Econometrica*, 61, 29–56.

Kim, Y. (1994): "Evolutionary Stable Strategies in the Repeated Prisoner's Dilemma," *Mathematical Social Sciences*, 28, 167–197.

Kuhn, H. W. (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games II,* Annals of Mathematics Study 28, ed. by H. W. Kuhn, and A. W. Tucker, pp. 193–216. Princeton University Press, Princeton NJ.

Lehrer, E. (1988): "Repeated Games with Stationary Bounded Recall Strategies," *Journal of Economic Theory*, 46, 130–144.

Nachbar, J. H. (1997): "Prediction, Optimization, and Learning in Repeated Games," *Econometrica*, 65, 275–309.

——— (2001): "Bayesian Learning in Repeated Games of Incomplete Information," *Social Choice and Welfare*, 18(2), 303–326.

——— (2005): "Beliefs in Repeated Games," *Econometrica*, 73, 459–480.

Osborne, M. J., and A. Rubinstein (1994): *A Course in Game Theory*. The MIT Press, Cambridge, Massachusetts.

Pearce, D. (1987): "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," Cowles Foundation Discussion Paper No. 855.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144, 379–396.

Rubinstein, A. (1979): "Equilibrium in Supergames with the Overtaking Criterion," *Journal of Economic Theory*, 21, 1–9.

Selten, R. (1983): "Evolutionary Stability in Extensive Two-Person Games," *Mathematical Social Sciences*, 5, 269–363.

Sugden, R. (1986): *The Economics of Rights, Cooperation and Welfare*. Basil Blackwell, Oxford.

van Damme, E. (1987): *Stability and Perfection of Nash Equilibria*. Springer Verlag, Berlin.

——— (1989): "Renegotiation-Proof Equilibria in Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 47, 206–217.

Young, H. P. (1993): "The Evolution of Conventions," *Econometrica*, 61, 57–84.