Evolution and Revolution: The Dynamics of Corruption

by

Cristina Bicchieri and Carlo Rovelli

May 1994

Report CMU-PHIL-57

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Evolution and Revolution: The Dynamics of Corruption

Cristina Bicchieri
Carnegie Mellon University
cb36@andrew.cmu.edu

Carlo Rovelli
University of Pittsburgh
rovelli@vms.cis.pitt.edu

May 1994

Abstract

In this paper we model the evolution of a system of corruption. We assume a fixed population of players that play a series of supergames with randomly chosen opponents. Each stage game in the supergames is a prisoner's dilemma. We show the conditions under which an equilibrium of corruption exists and is stable. We assume there are two types of players, adaptive and nonadaptive ones. Among the nonadaptive players, there is a small proportion that always chooses to be conditionally honest in every new supergame. Furthermore, we assume that corruption generates small but cumulative social costs. We show that the joint presence of a small group of "honest" players and of cumulative social costs is sufficient to drive the system to a critical (i.e., catastrophic) point in which the stable equilibrium of corruption suddenly becomes unstable. When the system has reached such a catastrophic point, a small perturbation is enough to drive it towards a different equilibrium. We show that the new equilibrium is cooperative, in that all players choose to be conditionally honest, and that a cooperative equilibrium is always stable under our model's conditions.

1. Introduction

Social norms have traditionally been linked with the problem of attaining social order or some form of social coordination. Social order can be guaranteed by a centralized authority, as in the classic Hobbesian view of Leviathan, or it can emerge spontaneously from the interaction of large numbers of individuals who did not plan nor expect it to occur. An early instance of such a spontaneous order, or "good anarchy", is Locke's state of nature. Perhaps the archetypal model of spontaneous social order is Adam Smith's description of the workings of the market, where the private pursuit of egoistic interests leads to an unintended, socially desirable outcome. In general, when we witness coordinated behavior that takes place without the monitoring and sanctioning intervention of a central authority, we tend to attribute its occurrence and persistence to the existence of social norms. Such norms prescribe socially adaptive behavior and/or proscribe behavior that is perceived as dangerous, antisocial, or plainly inappropriate.

When we refer to the power of norms in warranting social order, the regulation of conflict, or any other type of social coordination, we must refrain from the temptation of ascribing some necessarily positive social attribute to such an arrangement and to the norms that sustain it. An organization like the Mafia displays a remarkable degree of coordinated activity and its operations are regulated by norms (Gambetta 1993), but it would be hard to claim that its workings benefit society at large. Analogously, stable social systems in which discriminatory norms against women or minorities persist may turn out to be very inefficient from an economic standpoint (Elster 1989).

In this paper we study the norms that support a corrupt social system, i.e. a social system in which the illegal exchange of bribes is the norm. Norms may be defined in various ways, depending on the framework of analysis (Gibbs 1965, Cancian 1975). Our framework is a dynamic one, in that we study how certain norms evolve, stabilize or disappear. The goal of our investigation is to see under which conditions corruption survives, and what conditions favor its extinction and the emergence of a new system of norms. For our purposes, a behavioral definition of norms in terms of stable behavioral patterns is adequate.

It is important to note that corruption takes place in a framework characterized by exchanges, rather than by extortion and violence; though the exchange system is illegal, it is not enforced by threats nor is it regulated by some kind of centralized authority. The kind of corruption we shall model is the illegal exchange of bribes. Such exchange could be represented as a kind of informal cooperation among corrupt politicians and contractors that exchange bribes for contracts. We choose instead to model the noncooperative side of such exchanges. That is, we model the fact that a contractor/politician is always competing for scarce resources (contracts or bribes) against a group of other contractors/politicians. A contractor, for example, can be modeled as involved in a sequence of prisoner's dilemma games with other contractors. Such games are prisoner's dilemmas because, though it is individually better to be corrupt (i.e., to offer a bribe), the collective outcome of generalized corruption is much worse than the collective outcome of generalized honesty. Similar considerations apply to politicians. When a system of corruption is stable, it is a good example of ongoing spontaneous coordination. However, it is also a system that may involve huge inefficiencies.

An immediate question raised by any kind of protracted illegal exchange is how can this system of exchange -- and the norms that support it -- be stable and last for a long time without any apparent challenge. Another important question is what may cause the sudden and unexpected collapse of a generalized system of corruption, as well as the emergence of a new, different system. Such questions are common to all social norms. Indeed, one of the most important features of norms is that they can change rather quickly and, to some extent, unexpectedly. One would expect inefficient norms to disappear more rapidly and with greater frequency than more efficient ones. However, the fact that a norm may be inefficient, or that its being followed may generate social costs, is only a necessary condition for its demise. It is not a sufficient condition.

Though norms have been extensively studied in the social sciences and in psychology, very little attention has been devoted to understanding the dynamics of norms. Our analysis of the evolution of a system of corruption is meant to be an example of a more general view of how social norms evolve, spread or collapse. Since our starting point will be the analysis of a state in which almost everybody is corrupt, we do not analyze by which

mechanisms norms of corruption might emerge, or the conditions under which they are most likely to emerge.[1] Our analysis will be exclusively focused on the dynamics of certain behavioral patterns, and such dynamics will be modeled as an evolutionary process.

An evolutionary approach is based on the principle that strategies that make a person do better than others will be retained, while strategies that lead to failure will be abandoned. The success of a strategy is measured by its relative frequency in the population at any given time. To model individual choices, we use a game-theoretic framework. A repeated (finite) interaction with the same group of players is modeled as a supergame, and we assume that players play a series of supergames with randomly selected opponents. The payoff to an individual player depends on her choice as well as on the choices of the other players in the game. In an evolutionary model, however, players' choices are not strategic or fully rational, in that we do not assume that they want to maximize their expected utility over time (Nachbar 1990, Binmore and Samuelson 1992). We just assume that behavior is adaptive, so that a strategy that did work well in the past is retained, and one that fared poorly will be changed. The evolution of strategies is thus the consequence of adaptation. There are many different adaptive mechanisms we may attribute to the players. A realistic adaptive mechanism is learning by trial and error; another plausible mechanism is imitation: Those who do best are observed by others who subsequently emulate their behavior (Hardin 1982). An important advantage of the evolutionary approach is that it does not require sophisticated strategic reasoning in circumstances, such as large-group interactions, in which it would be unrealistic to assume it.

Strategies change over time as a function of their relative success in an environment that is made up of other players that keep changing their own strategies adaptively. There are several environments one may start with. A population of individuals can be represented as entirely homogeneous, in the sense that everybody is adopting the same type of behavior, or heterogeneous to various degrees. In the former case, it is important to know whether the commonly adopted behavior is stable against mutations. The relevant concept here is that of an *evolutionarily stable strategy* (Maynard Smith and

---

[1] For an analysis of how and under what circumstances norms might emerge, see J. Coleman (1989), K. Opp (1979, 1983), and E. Ullmann-Margalit (1977).

Price 1973; Taylor and Jonker 1978); when a population of individuals adopts such a strategy, it cannot be successfully invaded by isolated mutants, since the mutants will be at a disadvantage with respect to reproductive success. A more interesting case, and one relevant to a study of the reproduction of norms of corruption, is that of a population in which several competing strategies are present at any given time. What we want to know is whether the strategy frequencies that exist at a time are stable, or if there is a tendency for one strategy to become dominant over time.

There is a major difference between the model presented here and other evolutionary models of norms, such as Axelrod (1986) and Young (1993). Like Axelrod and Young, we want to show how certain patterns of behavior may evolve, and give the conditions under which they become stable. Unlike their models, however, our model focuses on the dynamics of change. In particular, we want to show how -- if the parameters of the model depend on an external variable and vary slowly -- a sudden change from stability to instability might occur. What we want to explain is precisely how norms that seem well established and almost permanent can suddenly collapse, and new norms get established in a relatively short span of time. The hypothesis we advance is that a progressive, slow accumulation of social costs may eventually lead to a catastrophe, i.e. to the sudden collapse of the entire system. The time evolution of norms that we describe is a typical example of a discontinuity emerging from a small variation of a continuous parameter (cumulative social costs, in our case), of the type studied by catastrophe theory. The interesting feature of the catastrophe theory approach is the fact that it relies on a qualitative analysis of the phenomenon, and that the necessity of the point of discontinuity -- the "catastrophic" event -- can be inferred from global properties of the system. Indeed, a property of the transition that we shall describe is that it is very stable under small modifications of the model.

It is important to note that in our model the progressive cumulation of social costs is not sufficient to induce a change in the system. Rather, a crucial role in the establishment of a new norm is played by a small percentage of "irreducibly honest" individuals. Such individuals are not adaptive in the sense that they never change their strategy. In particular, they always try to be conditionally honest when they start to play a new supergame. To be conditionally honest is equated with playing a tit-for-tat strategy. Remarkably

enough, in the presence of increasing social costs these infrequent attempts at honesty are sufficient to drive society towards the adoption of a new norm of honesty.

2. The Model

Corruption in this paper refers to the illegal exchange of bribes for contracts between politicians and contractors. There are many other kinds of corruption one may want to study, but for simplicity we limit our analysis to this case. Our conclusions are nevertheless very general, as they do not depend on the type of corruption that is being studied.

A generalized state of corruption can be represented in several ways. In particular, any form of corruption has both a cooperative and a noncooperative side. Even if in this paper we do not consider the cooperative aspects of corruption, let us just briefly mention the characteristics of such cooperative arrangements. Since bribes are illegal, transactions involving them would seem to be extremely costly at least in two respects. First, there is the risk of penal sanctions. Second, there is a constant risk that the other party does not fulfill his part of the illicit bargain. Each party is thus faced with a prisoner's dilemma in which he faces the choice of honoring the agreement or breaking the promise. The stability and pervasiveness of corruption can be explained by the fact that interactions are not a one-shot affair. Rather, they are repeated over time, either with the same partners or with different parties who know how one has behaved in the past. As is well known in the game-theoretic literature, the presence of reputation effects and the possibility of punishing defection by simply excluding the defector from future interactions are sufficient incentives for cooperation (Kreps et al. 1982, Kreps 1990). If the interaction between politicians and contractors is repeated over time, game theory predicts that - if certain conditions are fulfilled - there will be a cooperative equilibrium in which corruption is the rule. In such equilibrium the parties have an incentive to fulfill the illicit pacts because of their interest in the continuation of the relationship. A stable equilibrium of corruption is a norm (Bicchieri 1990, 1993). Everyone expects everyone else to offer a bribe or accept to be bribed, depending on the role they play in the exchange. And everyone prefers to conform to this pattern of behavior if everyone else conforms, too.

Our analysis will be centered instead on the noncooperative aspect of corruption. When a contractor has to decide whether to offer a bribe, what matters to him is the presence of competing contractors that may or may not behave honestly. If all his competitors are honest, it is to his advantage to offer a bribe. And if at least one of his competitors is corrupt, again it is better to bribe, in the hope of making an offer high enough to win the contract. Corruption is thus a dominant strategy. From this viewpoint, honest behavior is cooperative behavior, but the noncooperative option dominates. Similarly for the politician. Her competitors are other politicians that share the control of the same resources, and that may or may not ask or accept bribes. Accepting bribes puts a politician at an advantage (in terms of power and influence) over those competitors that did not accept bribes. Again, accepting a bribe is the dominant option. However, even if being corrupt is the best individual choice in a one-shot game, everyone would be better off in the long run if everybody refrained from offering/taking bribes. That is, collective honesty will fare better than collective corruption in any repeated game.

In our model we assume a fixed heterogeneous population P of agents that interact through small-group interactions in which they have to choose between honest and corrupt behaviour. The model is rather simple in that players only have two strategies to choose from, and we add the further simplification that the number of players that play any given supergame is constant. However, there would be nothing conceptual to gain in presenting a more general model, since the results we obtain in this simple case are easily generalizable to far more complex systems. The assumptions of the model are stated below.

1. We assume the individuals to interact in groups of n individuals, randomly chosen. We denote a protracted interaction between (the same) n individuals as a *supergame*. A supergame consists of N repetitions of a stage game between the n individuals. In our model of corruption, there will typically be a small group of contractors making tenders for public works to a small group of politicians that share the control of a given area. Each agent will have repeated interactions with the same group of politicians/contractors for some period of time. And each agent will also be involved in subsequent

new kinds of protracted interactions with different groups of politicians/contractors.

2. Each stage game is a prisoner's dilemma in which a player has the option of choosing to be honest (**h**) or corrupt (**c**). We assume each player to play against the group of (n-1) similar players. Corruption is the dominant strategy, in that playing **c** is better than playing **h**, whatever the opponents do. The payoff of each player depends on his own choice and on whether or not at least one other player is corrupt. In our example, in each stage game a player must choose whether to pay a bribe or not (or accept a bribe or not), where the opponents are agents that compete for the same resources. That is, the single contractor will have to decide whether to bribe or not, and his payoff will depend upon the honesty or corruption of the contractors that are his opponents in the stage game. Similarly the politician will have to decide whether to ask for a bribe or not, where her opponents in the stage game are other politicians that share the control of the same resources. Corruption is a dominant strategy, in that if all the other players are honest it pays to be corrupt, and if at least one of the opponents is corrupt it is better to be corrupt, too.

3. The players play a series of supergames (i.e., a series of N repetitions of the same stage game). After each round of play the actual payoffs and strategies of the players become public knowledge. Furthermore, the outcome of each supergame is known by all individuals in the population. In each supergame, a player can choose one among several strategies **S**. We consider (for simplicity) only two strategies for the supergame:
-- A "corrupt" strategy, which we denote as **C**, consisting in the choice of **c** in every stage game (constant defection);
-- A "tit-for-tat" strategy, which we denote as **H**, consisting in playing **h** in the first stage game of the supergame, and then playing **h** if all the opponents did play **h** in the previous stage game, and **c** if at least one opponent did play **c** in the previous stage game. (Note that the "honest" strategy **H** represents a form of conditional cooperation.)

4. A player can be one of two types. We assume that the large majority of players, which we call *strategic* players, choose a fixed strategy for the entire

supergame, but may readjust their strategy at the end of each supergame (or, equivalently, after an arbitrary number of supergames) according to its relative success. Thus a strategic player may be conditionally honest in one supergame and always corrupt in another. We interpret this slow change as the "sluggishness" of behavioral patterns. The second type of player is the "irreducible" one. Such player chooses a strategy once and for all, irrespectively of its relative success. An irreducible player may thus play tit-for-tat in all supergames, or instead he may always choose corruption. We assume the number of irreducible players, both honest and corrupt, to be very small and to remain the same through time. The larger part of the population will thus be composed of strategic players in varying proportions of tit-for-tatters and constant defectors.

5. The strategy of the strategic player is readjusted as follows. If p_{st} is the percentage of the strategic players that follow the strategy s in the supergame that starts at time t, and u_{st} is the expected payoff of playing the strategy s in this supergame, then at the end of the supergame p_{st} will change by getting multiplied by an arbitrary monotonically increasing positive function f of the expected payoff u_{st}. Thus, the readjustment yields

$$p_{s\,t+1} = Z\, f(u_{st})\, p_{st}\,, \qquad (1)$$

where we have taken the time in which one supergame is played to be the time unit, and Z is a normalization factor, independent of s, which renormalizes the probabilities so that their sum at t+1 remains equal to one.
6. Finally, we assume that the payoffs of the stage game change slowly in time. In particular, we assume that there is a slow decrease in time of all the payoffs, which we denote as *payoff erosion*. However, the stage game *remains a prisoner's dilemma.* The slow change in the payoff values is meant to reflect the accruing social cost of corruption, in terms of inefficiencies and wasted resources.

Our aim is to study the evolution of the strategies, namely the change in the proportions of individuals choosing the different strategies, assuming that the number of individuals and the number of supergames is large.

3. Analysis of the model

Each player faces the following prisoner's dilemma matrix:

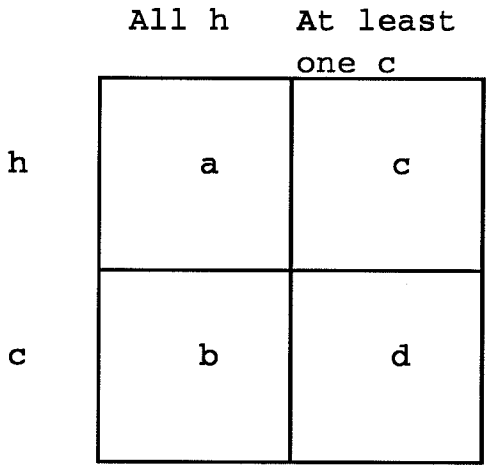

Figure 1: Stage-game payoffs

The letters appearing in the matrix of Figure 1 represent the payoffs obtained by the row player for each combination of his and the opponents' strategies. Because we assume the game to be symmetrical, each player faces the same matrix, and since the stage game is a prisoner's dilemma (PD), we have that $b > a > d > c$. We normalize the payoffs by taking $c = 0$. The best collective outcome results from universal honesty (each player gets a), but the best individual choice is to be corrupt (one gets b if all other players are honest, d if at least one is corrupt).

We model the slow payoff erosion by a joint decrease of a, b and d by a small amount ε at every supergame. Since we have taken $c = 0$, we assume that ε is small enough to keep d always positive, so that the stage game remains a prisoner's dilemma. Note that c stays equal to 0 since the honest individual, when matched with a group in which at least one player is corrupt, does not gain nor lose anything. We start our analysis by studying the evolution of the system without payoff erosion; later we consider the effect of payoff erosion.

We assume that a player's total payoff in a supergame is the undiscounted sum of the payoffs she gets at each stage game. What follows is

the total payoff matrix of the row player for N repetitions of the stage game of Figure 1. Note that we have limited the number of possible strategies in the supergame to tit-for-tat (**H**) and always corrupt (**C**). Recall that tit-for-tat is the choice of being conditionally honest. Such player will play honest in the first stage of the supergame, and subsequently will choose to be honest/corrupt depending on the behavior of the other players in the previous stage game. To be always corrupt instead means that a player will choose to be consistently corrupt for the entire duration of the supergame.

	All H	At least one C
H	Na	(N-1)d
C	(N-1)d+b	Nd

Figure 2: Supergame payoffs

Let M be the (small) number of players that follow a fixed strategy in all supergames, out of which m_H are "irreducibly honest" players who always play the **H** strategy, whereas m_C are the "irreducibly dishonest" players, who always play the **C** strategy.

Let $N >> M$ be the number of "strategic" players that readjust their strategy at each supergame. Let $P = N + M$ be the total number of players in the population. As before, we denote by n the number of players that play together in a given supergame. For simplicity, we assume such number to remain constant. At any given time, there will be a number of supergames being played by the population P, each supergame being played by just n players.

Let n_{Ht} (respectively n_{Ct}), be the number of strategic players that play **H** (respectively **C**) at time t, but are ready to readjust their strategy in the next supergame. Let us define the percentages (or relative frequencies) of the various strategies as:

$\pi_H = m_H/P$ irreducibly honest,
$\pi_C = m_C/P$ irreducibly corrupt,

$p_{Ht} = n_{Ht}/P$ strategically honest,
$p_{Ct} = n_{Ct}/P$ strategically corrupt.

In a supergame played with randomly chosen opponents, the probability of playing against n-1 tit-for-tatters (**All H**) is $(p_{Ht} + \pi_H)^{n-1}$, and the probability of being matched with at least one corrupt individual is $1 - (p_{Ht} + \pi_H)^{n-1}$. Then in any supergame the expected payoffs of the strategies **H** and **C** are given by the sum of the total payoffs a player obtains by playing against different opponents' strategies (see Figure 2), weighted by the probability of being matched with any such strategies. That is, using Figure 2:

$$u_{Ht} = Na\,(p_{Ht} + \pi_H)^{n-1} + (N-1)d\,[1-(p_{Ht} + \pi_H)^{n-1}], \quad (2)$$
$$u_{Ct} = [(N-1)d+b]\,(p_{Ht} + \pi_H)^{n-1} + Nd\,[1-(p_{Ht} + \pi_H)^{n-1}].$$

At the end of each supergame, we have a readjustment of strategies. Following equation (1), the new normalised frequency p_H then becomes

$$p_{H\,t+1} = Z\,f(u_{Ht})\,p_{Ht} \quad (3)$$

Similarly, the new normalised frequency p_C becomes

$$p_{C\,t+1} = Z\,f(u_{Ct})\,p_{Ct} \quad (3')$$

where Z is the normalization factor. Z is determined by the requirement that the sum of the probabilities must be 1. Note that, because of the irreducibles in the population, we must have that $p_H + p_C = N/P$. Therefore we must have that

$$Z = N/P \quad \frac{1}{f(u_{Ht})\, p_{Ht} + f(u_{Ct})\, p_{Ct}} \tag{4}$$

Since the only two variables are p_{Ht} and p_{Ct}, and their sum is always N/P, the configuration of the system is completely determined by the percentage p_{Ht}. Thus, what we are concerned with is the time evolution of p_{Ht} determined by equations (3), (4) and (2). From now on, we use (for simplicity) p_t for p_{Ht}, and $(N/P) - p_t$ for p_{Ct}. Thus, our main evolution equation is (from (3) and (4)):

$$p_{t+1}(p_t) = N/P \quad \frac{f(u_{Ht}(p_t))\, p_t}{f(u_{Ht}(p_t))\, p_t + f(u_{Ct}(p_t))\, (N/P - p_t)} \tag{5}$$

where $u_{Ht}(p_t)$ and $u_{Ct}(p_t)$ are, from equation (2),

$$\begin{aligned} u_{Ht}(p_t) &= Na\,(p_t + \pi_H)^{n-1} + (N-1)d\,[1 - (p_t + \pi_H)^{n-1}], \\ u_{Ct}(p_t) &= [(N-1)\,d + b]\,(p_t + \pi_H)^{n-1} + Nd\,[1 - (p_t + \pi_H)^{n-1}]. \end{aligned} \tag{6}$$

4. Equilibria

Let us now study the conditions at which the system is in equilibrium. Clearly, the equilibrium is characterised by $p_{t+1} = p_t$. In this case, there is no change in the relative frequencies of the different strategies from one period to another. From equation (5), the equilibrium is characterised by

$$p_t = N/P \quad \frac{f(u_{Ht}(p_t))\, p_t}{f(u_{Ht}(p_t))\, p_t + f(u_{Ct}(p_t))\, (N/P - p_t)}$$

From this equation, we have

$$N/P \quad f(u_{Ht}(p_t))\, p_t - [\, f(u_{Ht}(p_t))p_t + f(u_{Ct}(p_t))(N/P - p_t)\,]\, p_t = 0,$$

or

$$p_t\, [f(u_{Ht}(p_t))\, (\, p_t - N/P \,) - f(u_{Ct}(p_t))\, (\, p_t - N/P \,)\,] = 0.$$

Namely

$$p_t \quad [f(u_{Ht}(p_t)) - f(u_{Ct}(p_t))] \quad [p_t - N/P \,] = 0. \tag{7}$$

This is the condition for the system to be in equilibrium. In other words, the equilibria of the systems are given by the values of p_t included between 0 and N/P that satisfy equation (7). Again, an equilibrium is characterized as a situation in which there is no change in the proportion of players who play any strategy. Equation (7) is satisfied in the following three cases:

i. $p_t = 0$

ii. $p_t = N/P$

iii. $f(u_{Ht}(p_t)) = f(u_{Ct}(p_t))$; since the function f is monotonically increasing, this implies:
$u_{Ht}(p_t) = u_{Ct}(p_t)$.

Accordingly, there can be three equilibrium configurations in the system, which we denote as (i), (ii) and (iii). Let us start by considering case (i), the case in which $p_t = 0$. This means that the proportion of strategic players that choose to play **H** in any supergame at time t is zero. This is an equilibrium in which the population of strategic players remains corrupt in every supergame.

Case (ii) is a situation in which all the strategic players choose to be conditionally honest (*i.e.* play tit-for-tat) in any supergame starting at time t. Note that the presence of a small number of irreducibly corrupt agents does not change the result. Since their number is very small, the number of supergames in which the strategic players will end up choosing corruption is extremely small.

The third equilibrium is a state in which the expected payoff of conditional honesty is equal to the expected payoff of corruption for all supergames at time t. Explicitly, we have, using equation (6)

$$\begin{aligned} &Na\,(p_t + \pi_H)^{n-1} + (N-1)d\,(1-(p_t + \pi_H)^{n-1}) \\ &- [(N-1)\,d+b]\,(p_t + \pi_H)^{n-1} - Nd\,(1-(p_t + \pi_H)^{n-1}) = 0 \end{aligned}$$

or

$$(p_t + \pi_H)^{n-1}\,[\,Na-(N-1)d - [(N-1)\,d+b]+Nd] = -(N-1)d + Nd = d,$$

or

$$(p_t + \pi_H)^{n-1} = \frac{d}{Na-Nd+d-Nd+d-b+Nd}.$$

which gives

$$p_t = \sqrt[n-1]{\frac{d}{N(a-d)+2d-b}} - \pi_H \qquad (8)$$

It is important to notice that p_t must be always greater or equal to zero. (If p_t is equal to zero, we have case (i) again.) If, on the other hand, the right-hand side of equation (8) is less than zero, there is no equilibrium corresponding to case (iii). Thus an (independent) equilibrium (iii) exists only if p_t is greater than zero, in which case it must be true that

$$\pi_H{}^{n-1} < \frac{d}{N(a-d)+2d-b} \cdot \qquad (9)$$

Thus, we have two distinct possibilities. If equation (9) is satisfied, then there is an equilibrium point corresponding to case (iii), and thus we have the three equilibria (i), (ii), and (iii). Intuitively, this happens if there are few irreducibly honest players in the population. We denote this situation as the *first regime.* On the contrary, if equation (9) is not satisfied, we only have the two equilibria (i) and (ii). In this case, enough (or too many) irreducibly honest players are present in the population. We call this situation the *second regime.*

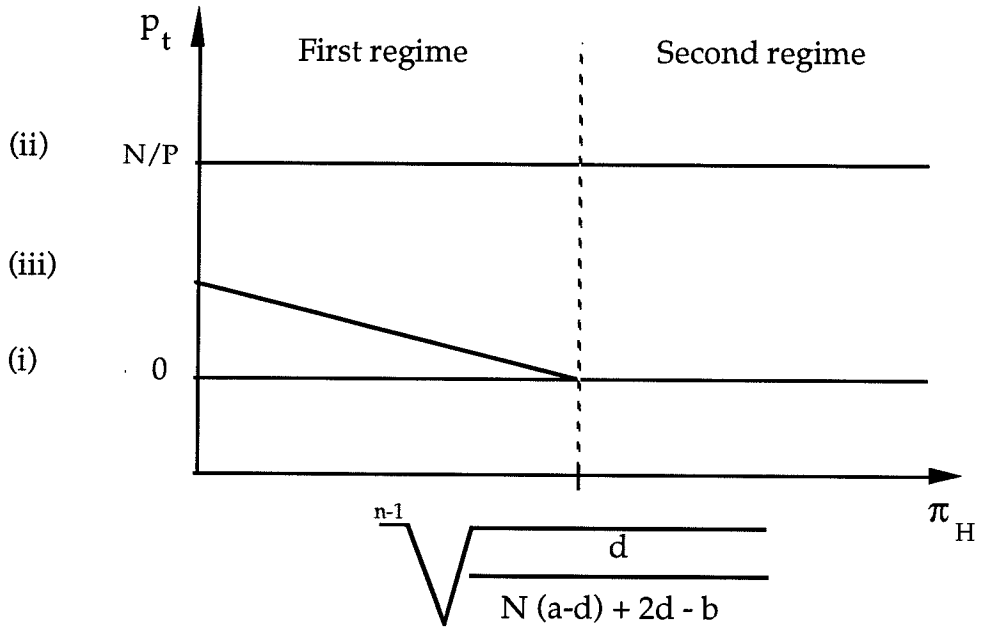

Figure 3: The equilibrium states for different values of π_H

5. **Stability**

Next, let us analyze the stability properties of the equilibrium points. The condition for an equilibrium point to be stable is that a small increase in p_t leads to a decrease in p_{t+1}, and vice versa, a small decrease in p_t leads to an increase in p_{t+1}. Consider the function $p_{t+1}(p_t)$, defined by equation (5). Since the equilibrium points are given by the values of p_t for which $p_{t+1}(p_t) = p_t$, we can represent graphically the equilibria as the intersections between the graph of the function $p_{t+1}(p_t)$ and the straight line $p_{t+1} = p_t$. (See Figure 4)

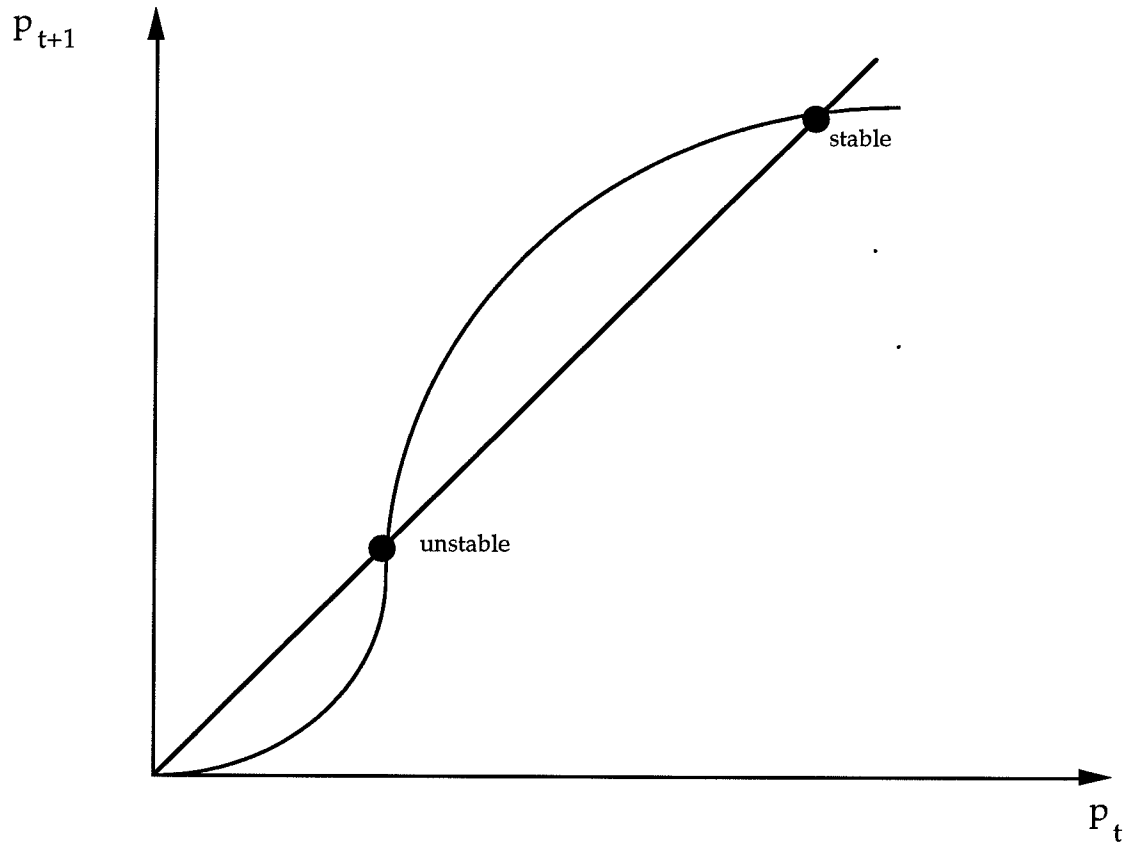

Figure 4. Equilibria and stability

There are two kinds of such intersections, the ones from left to right and the ones from right to left. In the former we have

$$\frac{dp_{t+1}}{dp_t} < 1. \tag{10}$$

In the latter we have $dp_{t+1}/dp_t > 1$. By looking at Figure 4, it is easy to see that the first kind of intersection, where equation (10) is satisfied, is one in which a small increase in p_t leads to a decrease in p_{t+1}, and a small decrease in p_t leads to an increase in p_{t+1}. Thus equation (10) represents the stability condition for an equilibrium. Let us now compute the derivative (10) at the equilibrium points. In order to simplify the expressions in the computation, let us put

$$A = f(u_{Ht}(p_t)) ,$$

$$B = f(u_{Ct}(p_t)) ,$$

and let us denote with a prime (') the derivation with respect to p_t. A simple derivation yields

$$\frac{dp_{t+1}}{dp_t} = N/P \quad \frac{(A + A'p_t)\,[Ap_t + B(N/P - p_t)] - Ap_t\,[(A + A'p_t)+B'(N/P - p_t) - B]}{[Ap_t + B(N/P - p_t)]^2}$$

Thus, the stability condition (10) becomes

$$N/P \quad \left[(A + A'p_t)\, B\, (N/P - p_t) - Ap_t\,[B'(N/P - p_t) - B]\right] < [Ap_t + B(N/P - p_t)]^2 \qquad (11)$$

Let us now consider the three equilibrium points (i), (ii) and (iii) separately. The first one ($p_t = 0$) is stable if equation (11) is satisfied when we put $p_t = 0$. This gives

$$N/P \quad [A\, BN/P\ ] < [B\ N/P\ ]^2$$

Since N, P and B are positive numbers, we have that

$$A < B$$

namely $f(u_{Ht}(p_t)) < f(u_{Ct}(p_t))$. Since f is monotonically increasing, this implies that $u_{Ht}(p_t) < u_{Ct}(p_t)$. An equilibrium in which all the strategic players choose strategy **C** is stable as long as the expected payoff of corruption is greater than the expected payoff of conditional honesty (**H**). Using equation (6) (with $p_t = 0$), the stability condition is

$$Na\ \pi_H^{n-1} + (N-1)d\ [1 - \pi_H^{n-1}] < [(N-1)d+b]\ \pi_H^{n-1} + Nd\ [1 - \pi_H^{n-1}].$$

From this, we obtain

$$\pi_H^{n-1}\ [Na - 2Nd + 2d - b + Nd] < d\ .$$

or

$$\pi_H^{n-1} < \frac{d}{N(a - d) + 2d - b} \quad . \qquad (12)$$

This is the condition for the stability of the equilibrium (i). This means that all strategic players will keep playing **C** if the number of irreducibly honest players is less than a certain value.

Now, notice that equation (12) is the same condition as equation (9) (which is the condition for the existence of equilibrium (iii), or for being in the first regime). Thus, we can conclude that *the equilibrium (i) is stable in the first regime, but becomes instable in the second regime.*

The stability properties of the other two equilibria follow from the stability of (i) by elementary analysis theorems: Since the equilibria are determined by the intersections between the graph of the function $p_{t+1} = p_{t+1}(p_t)$ and the straight line $p_{t+1} = p_t$, and an equilibrium is stable if the first function intersects the second from left to right, it follows immediately that the equilibria are alternatively stable and unstable (see Figure 4). Thus we have the following situation:

Equilibria	**i**	**iii**	**ii**
first regime	stable	unstable	stable
second regime	unstable	non existent	stable

Notice that the equilibrium (ii) in which $p_t = N/P$, i.e. all strategic players chose to be conditionally honest, is always stable.

6. **Payoff erosion**

Up to now, we have neglected the effects of payoff erosion. Let us now consider its effects on the evolution of the system. The only change from the previous model is that the stage game payoffs a, b and d evolve in time (recall that c = 0). Since the payoffs decrease by a very small amount ε at every supergame, i.e., at every unit of time t, at any time t they will be given by

$a_t = a - \varepsilon t,$
$b_t = b - \varepsilon t,$
$d_t = d - \varepsilon t.$

We have assumed ε to be small enough so that we are always in a prisoner's dilemma situation, that is $b_t > a_t > d_t > 0$. Thus, we must have

$$\varepsilon < d/T, \tag{13}$$

where T is the maximum time we consider. T can be tought of as the threshold after which d_t becomes negative. Then, the analysis of the model we have provided remains valid at all times between 0 and T. However, equation (9), which separates the first regime from the second regime, is affected by the evolution of the system. Specifically, we have that the system is in the first regime at time t if

$$\pi_H{}^{n-1} < \frac{d_t}{N(a_t - d_t) + 2d_t - b_t} .$$

That is

$$\pi_H{}^{n-1} < \frac{d - \varepsilon t}{N(a-d) + 2d - b - \varepsilon t} .$$

This condition is a function of time. Therefore there will be a time interval in which the system is in the first regime and another time interval in which it will be in the second regime. Specifically, solving with respect to t, the condion for being in the first regime becomes

$$t < \frac{d - [N(a-d)+2d-b]\,\pi_H{}^{n-1}}{(1 - \pi_H{}^{n-1})\,\varepsilon} = t_{critical} .$$

For $t < t_{critical}$ the system is in the first regime, for $t > t_{critical}$ the system is in the second regime. If the number of repetitions N is large, the term (2d-b) in the numerator is negligible compared to the term N(a-d); similarly, if the proportion of irreducibly honest players π_H is small (but not vanishing), the term $\pi_H{}^{n-1}$ in the denominator is negligible compared to 1. In this case the expression for $t_{critical}$ can be written in the simpler form

$$t_{critical} \sim \frac{d - N(a-d)\pi_H^{n-1}}{\varepsilon}.$$

7. Revolution

Let us consider the evolution of a society of individuals that start off at t = 0 in a state in which all the individuals follow the "Corrupt" strategy **C** (except for the small number of irreducible individuals that, at every new supergame, try once to play honest). At t = 0, we thus have $p_t = 0$. Let us assume that at t = 0 the system is in the first regime. Since in this regime the equilibrium (i), in which $p_t = 0$, is stable, the system will remain in such a state. In other words, the strategic players will find no incentive to deviate from the "corrupt" strategy. The system will remain in this state up to the time $t_{critical}$. When $t_{critical}$ is reached, the corrupt strategy is still dominant in the stage game, but the expected payoff of **H** in any supergame now exceeds the expected payoff of **C**. At this time the system enters into the second regime. The point $p_t = 0$ remains an equilibrium state, but it becomes unstable. Instability means that isolated clusters of cooperation can now trigger a cascade of honest behavior. Even if there is only a small probability that a group of irreducibly honest individuals cluster together in a supergame, if this happens they will do extremely well. Their behavior and payoffs can be observed by all other players, that will be driven to imitate the successful behavior.

Suddenly, at time $t_{critical}$ the society is driven toward the (only) other equilibrium, which is stable, namely $p_t = N/P$, in which all the strategic individuals chose the "honest" strategy **H**. The system has entered a phase in which the expected payoff for the cooperative "honest" strategy has overtaken the expected payoff of the corrupt strategy. This induces an increase in the number of players that choose the "honest" strategy, which in turn reinforces the advantages of cooperative behavior. The system stabilizes on a new equilibrium point (ii), in which all the strategic players choose to be conditionally honest. A new cooperative social norm thus gets established. The society has moved from the *first regime* to the *second regime*. From now on, to follow this new norm of honesty is to everybody's advantage. We denote this catastrophic event the "honesty revolution".

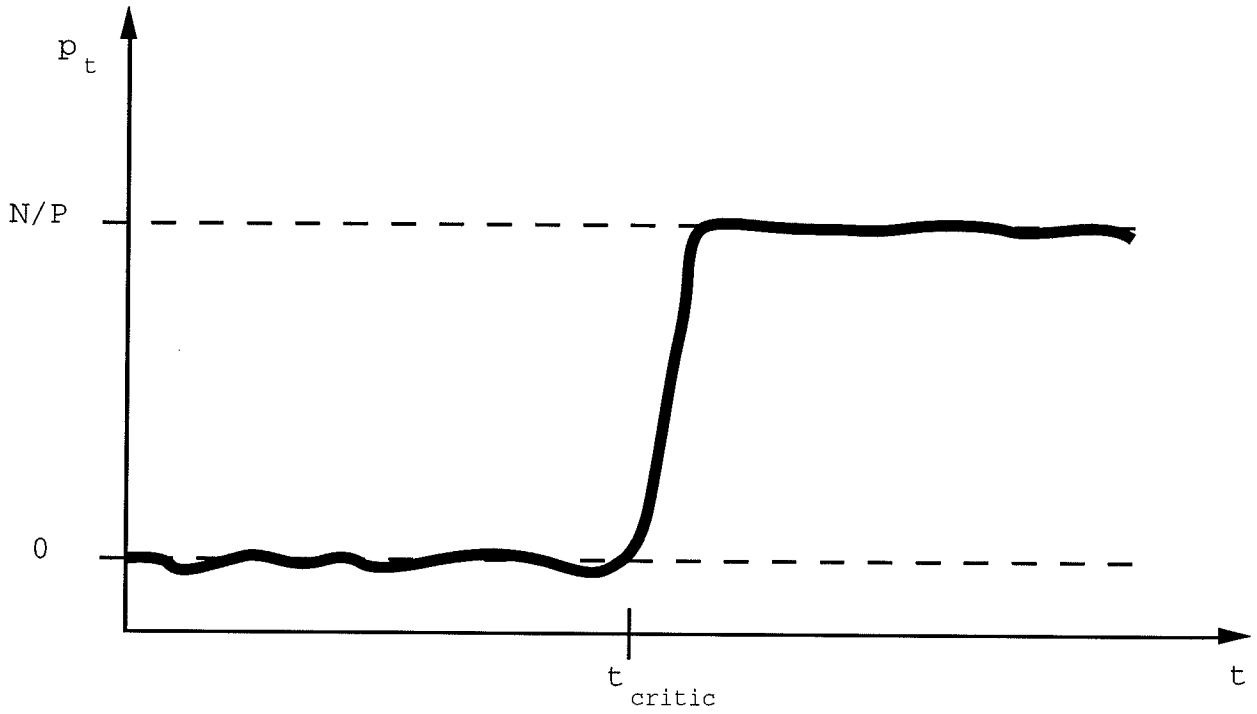

Figure 5: Time evolution of the
percentage of "honest" stategic players p_t

Figure 5 shows the typical evolution of a system that starts in a stable equilibrium. Note that since equilibrium (iii) is not stable, we have omitted it. The interesting change we have considered is the sudden, "catastrophic" change from stability to instability of an equilibrium system. We know from our model that equilibrium (ii), $p_t = N/P$, is always stable. The crucial case is thus the equilibrium $p_t = 0$, i.e. the equilibrium in which every strategic player chooses to be corrupt. This equilibrium remains stable up to a critical time. After that time, the system moves to a new equilibrium, in this case a cooperative one. When an equilibrium is stable, there are small random fluctuations around the equilibrium, which we represent in Figure 5 with an ondulated line. The intuitive interpretation of such fluctuations is that, in a stable corrupt system, there can be supergames in which some strategic players chose to play **H**. However, the evolutionary dynamics brings them back to playing **C**. Similarly, when the system has moved to the new

equilibrium in which the population of strategic players plays **H**, there can be fluctuations around this equilibrium.

We can visualise the "catastrophic surface" that describes the sudden change from stability to instability as follows (see Figure 6). Consider a function $F(p_t, t)$ that describes a surface over the (p_t, t) plane. For every fixed t, we choose $F(p_t)$ as a function that has minima in the points that represent stable equilibria. More precisely, we assume that for $t < t_{critical}$ the function F has minima in $p_t = N/P$, and $p_t = 0$ (stable equilibria), and a maximum (unstable equilibrium) in the point corresponding to the equilibrium (iii) (which lies between 0 and N/P). As t approaches $t_{critical}$, the equilibrium (iii) approaches 0 and, for $t > t_{critical}$, the function F has only one minimum in $p_t = N/P$. Since we are not interested in what happens around $p_t = N/P$, but only in what happens around $p_t = 0$, we restrict our attention to a strip of the (p_t, t) plane that includes $p_t = 0$, but not $p_t = N/P$. The function we have described defines a surface in the three-dimensional space (F, p_t, t). Consider the projection of this surface on the plane (F, t). For $t > t_{critical}$ this projection is injective, but for $t < t_{critical}$, there is a region R in the (F, t) plane in which the projection is not injective. Thus, we have a discontinuous projection of a continuous surface, namely a catastrophe, in the language of catastrophe theory. More precisely, the inverse image of a point in R includes three points of the surface. The region R is bounded by two lines, which meet at a point P, over $t_{critical}$. This is the point where the catastrophe occurs. This particular form of catastrophe is denoted as a cusp catastrophe.

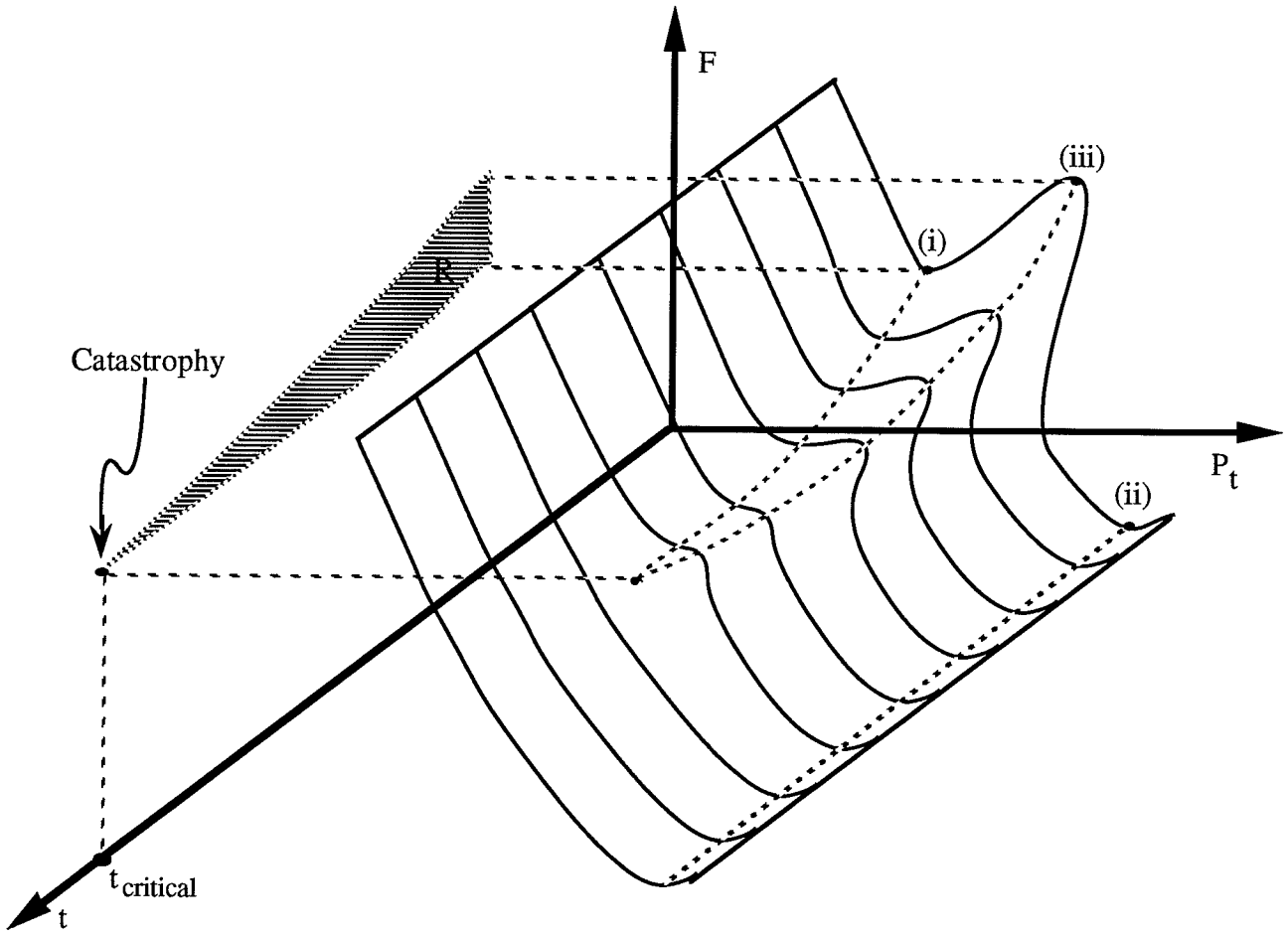

Figure 6: The catastrophic surface

8. **Conclusions**

We have modeled the evolution of a system in which there are two types of players that play a sequence of repeated prisoner's dilemma games with randomly chosen opponents. Each player has two possible strategies, and the large majority of players (the strategic ones) plays adaptively: When a strategy does not perform too well in a supergame, the player will choose another strategy in the next supergame. We gave the conditions under which the system is in equilibrium, as well as the stability conditions for such

equilibria. The most surprising result of our model is that cooperative behavior can suddenly emerge out of a system in which defection (corruption) is the norm. In particular, we show that -- provided that corruption generates small but cumulative social costs and the population contains a small number of irreducibly honest individuals -- there will come a critical time at which a stable equilibrium of corruption becomes suddenly unstable. This catastrophic event should be rightly called the "honesty revolution". An interesting feature of such a catastrophic change is that it is not simply caused by the social costs generated by the established norm. Only the combination of cumulative social costs and a small group of nonadaptive honest players is sufficient to eventually produce a discontinuity.

Note the essential role played by the small percentage of irreducibly honest individuals π_H. Since t never reaches the value d/ε (because of equation (12)) if $\pi_H{}^{n-1}$ is zero, namely if there are no irreducibly honest players, the system will never reach the critical time, and there will be no revolution. Thus, the presence of a small percentage of irreducibly honest individuals is essential in order to drive the honesty revolution. This percentage, however, can be extremely small. Note also that since the strategy that the irreducibly honest individuals play is tit-for-tat (conditional honesty), these individuals will choose the honest option **h** only once at the beginning of a supergame, unless they find out that all their opponents are behaving honestly, too. In the long run, these minor attempts are sufficient to drive society towards adopting a new norm of honesty, provided of course that the cumulative social costs of corruption progressively lower the advantages of corrupt behavior.

References

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

Axelrod, R. (1986) "An Evolutionary Approach to Norms", *American Political Science Review* 80: 1095-1111.

Bicchieri, C. (forthcoming) "Learning to Cooperate" in C. Bicchieri, and B. Skyrms (eds.),*Evolution, Learning, and Dynamics in Games,* Cambridge University Press.

Bicchieri, C. (1993) *Rationality and Coordination,* Cambridge University Press.

Bicchieri, C. (1990) ""Norms of Cooperation", *Ethics* 100: 838-861.

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

Cancian, F. (1975) *What are Norms?*, Cambridge University Press.

Coleman, J. (1989) *Foundations of Social Theory.*, Harvard University Press.

Elster, J. (1989) *The Cement of Society*, Cambridge University Press.

Foster, D. and P. Young (1990) "Stochastic Evolutionary Game Dynamics", *Theoretical Population Biology* 38: 219- 32

Gambetta, D. (1993) *The Sicilian Mafia*, Harvard University Press.

Gibbs, J. (1965) "Norms: The Problem of Definition and Classification". *American Journal of Sociology* 70: 586-594.

Hardin, R. (1982) *Collective Action*, The John Hopkins University Press.

Kreps, D. (1990) *Game Theory and Economic Modeling*, Oxford University Press.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982) " Rational Cooperation in the Finitely Repeated Prisoner's Dilemma" *Journal of Economic Theory* 27: 245-252.

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

Maynard Smith, J. and G. Price (1973) "The logic of animal conflict",*Nature* 246: 15-18.

Nachbar, J. (1990) "Evolutionary Selection Dynamics in Games: Convergence and Limit Properties",*International Journal of Game Theory* 19: 59- 89

Opp, K. (1979) "Emergence and Effects of Social Norms",*Kylos* 32: 775-801.

Opp, K. (1983) "Evolutionary Emergence of Norms",*British Journal of Social Psychology* 21: 139-49.

Taylor, P. and L. Jonker (1978) "Evolutionarily Stable Strategies and Game Dynamics", *Mathematical Biosciences* 40: 145- 156.

Ullmann-Margalit, E. (1977) *The Emergence of Norms*, Oxford University Press.

Young, P. and D. Foster (1991) "Cooperation in the Short and in the Long Run", *Games and Economic Behavior* 3: 145-156.

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