CentER Discussion paper

for
Economic Research

[redacted]

TILBURG UNIVERSITY

CentER
for
Economic Research

No. 9339

Competition or Co-operation

by Werner Güth
and Hartmut Kliemt

June 1993

ISSN 0924-7815

COMPETITION OR CO-OPERATION

Werner Güth*
Department of Economics
University of Frankfurt**

and

Hartmut Kliemt
Department of Philosophy
University of Duisburg***

Abstract

In the basic game of trust a first mover can decide between competition or trust in the other's willingness to respond in kind. Whereas the game ends in case of competition, the second mover can react to trustful cooperation by exploiting the first mover or by dividing the rewards evenly. Whereas the game theoretic solution predicts competition, one evolutionary analysis proves the evolution of a sufficiently strong conscience guaranteeing cooperation. If, however, the feeling of guilt is private information, the only evolutionarily stable result implies a non-preventive conscience and correspondingly an inefficient payoff vector. We finally discuss the possibility of investing in a perfect detection technology where two evolutionary stable situations can coexist, one relying on competition as well as non-monomorphic composition of the population.

---

* The author is indebted to CentER for its support and excellent research environment. The generous Alexander von Humboldt-award via the Dutch Science Foundation (NWO) which made his stay at CentER possible is gratefully acknowledged.

** P.O. Box 11 13 32 (Fach 81), D-6000 Frankfurt/M., Federal Republic of Germany.

*** Lotharstrasse 65, D-4100 Duisburg, Federal Republic of Germany.

Werner Güth and Hartmut Kliemt

Competition or Co-operation

On the evolutionary economics of trust, exploitation and moral attitudes

> "For he that performeth first, has no assurance the other will performe after; because the bonds of words are too weak to bridle mens ambition, avarice, anger, and other Passions, without the feare of some coercive Power; which in the condition of meer Nature, where all men are equall, and judges of the justness of their own fears cannot possibly be supposed. And therefore he which performeth first, does but betray himselfe ..." (Hobbes, Leviathan, chap. 14)

Human beings are endowed with foresight and understanding. They can form models of the world and anticipate future causal effects of their actions. Knowing that their fellows command the same faculties of the mind humans can behave strategically rational. On the other hand, emotions and passions may interfere with the precepts of reason and thus harm long run interests. Classical philosophers have discussed this quite extensively as "conflict between reason and passion". More recently economists revived the old philosophical insight that "harmful" emotions can be "checked" by beneficial ones. They pointed out that the emotions may adopt a strategic role that can further rather than harm individuals´ long run interests (cf. Frank 1987/88). In particular they can serve as "guarantors of promises" (cf. Hirshleifer 1987) and thus render subgame perfect (cf. Selten 1965, 1975) equilibria of strategic interaction which could not be reached by forward looking (incrementally) rational choices.

Subsequently we shall focus on an elementary but fundamentally important class of social interactions to which we shall refer as "games of trust". In such situations a second moving individual can exploit a first mover´s trustful co-operative behavior. If the first mover must suspect that the second mover behaves "competitively" rather than co-

operatively in games of trust the first mover shall not trust and a Pareto-inefficient result shall emerge. But mutually beneficial co-operation and a Pareto-efficient result become viable if first moving individuals can rationally expect that second movers are endowed with sufficiently strong moral attitudes like retributive emotions supporting a "sense of justice or fairness".

Describing this strategic role or function of emotions in furthering co-operation is of considerable interest in itself (cf. for a philosophical account of the decisive role of retributive emotions within any viable system of morals Mackie 1985). But it does not explain how the emotions could evolve and can prevail among human beings. Quite to the contrary competition between biological organisms in the Hobbesian jungle does not seem to leave clearance for the evolution of moral attitudes or, for that matter, a conscience. Indeed, ever since Darwin there was a general suspicion that the struggle for reproductive success must drive out or, more precisely, must have driven out any non-competitive dispositions. Philosophers like, most prominently Nietzsche, wondered how such a phenomenon like a conscience could emerge and survive in inter-individual competition. Even nowadays and regardless of our knowledge of such concepts like kin selection (cf. Maynard Smith 1964 and with respect to humans Alexander 1978) it may seem quite obvious -- at least at first sight -- that a monomorphic population of trustworthy individuals cannot be evolutionarily stable since it can be successfully invaded by a non-trustworthy mutant.

According to this view it should be expected that human individuals are not naturally endowed with a preference for fair retribution of trustful co-operative acts of their fellows. Biological competition should have "programmed" human actors to behave competitively in the sense of exploiting trustful behavior of their fellow humans. Social co-operation should not be viable unless some external coercive power interferes. However, such a Hobbesian view of social co-operation not only leaves us with the so-called "Hobbesian problem of social order" (cf. Parsons 1968) - in particular the problem of how to explain co-operation in

creating the external coercive power itself - it is also not in line with our general experience (cf. also the classical statement in Kropotkin 1902). For, evidently, besides those who exploit others there are at least some individuals who are at least sometimes trustworthy and in fact are willing to retribute fairly even though they could act differently.

In this paper we shall try to reach a somewhat deeper understanding of how and to what extent moral attitudes supporting co-operative behavior in "games of trust" might conceivably have evolved. In particular we address the issue of evolutionary stability of retributive emotions facilitating co-operation among rational actors. First, some elementary concepts of evolutionary game theory are sketched (1.). Then the "game of trust" is formally introduced (2.). In a third step this two person extensive game is analyzed from an evolutionary point of view (cf. for former applications of basically the same methods to other social settings Güth 1991, Güth and Yaari 1992). It is shown that retributive emotions can prevail in evolutionarily stable ways if players know beforehand whether or not the other player is endowed with a "(sufficiently strong) conscience" (3.1). On the other hand developing a "(sufficiently strong) conscience" shall not be evolutionarily stable if individuals, though knowing the distribution of types in the population, cannot identify the type of the other player before playing the game of trust (3.2). Afterwards our analysis of the foregoing two polar or extreme cases is extended to an intermediate case in which specific information about the type of the other player is available at a cost (3.3.). Some final observations and a general discussion of results conclude the paper (4.).

1. Basic Concepts of Evolutionary Theory

To put it very succinctly theories of biological evolution study the competition between alternative "genetic endowments" in a gene pool. Success is measured in terms of the relative frequency of genes. The so called "replicator dynamics" which determine relative frequencies in the

process of biological reproduction are quite well understood and can be described by "replicator equations".

More specifically, assume that M is a one dimensional "mutant space" – for convenience a subset of the real numbers $\Re$ – and $m \in M$ some mutant. Let the density function f_t: M —> [0, 1] characterize for each $t \in T$ the relative frequencies for all mutants $m \in M$ by $f_t(m)$. If the mutant m is matched with a competitor $m^\circ \in M$ the relative fitness of m against m° is measured by $H(m, m^\circ)$. The equation

$$(2.1)\ H_t(m) = \int_M H(m, m^\circ)\, f_t(m^\circ)\, dm^\circ$$

characterizes the relative reproductive success of $m \in M$ (reproductive success related to all $m^\circ \in M$ and their frequency distribution in the gene pool); while

$$(2.2)\ H_t = \int_M H_t(m)\, f_t(m)\, dm$$

is the average reproductive success of all mutants in period t.

According to the replicator equation

$$(2.3)\ \dot{f}_t(m) = f_t(m)\, [H_t(m) - H_t],$$

which complies with the requirement $\int_M \dot{f}_t(m)\, dm = 0$ for transforming one density function into another one, the mutant $m \in M$ can be relatively more successful in a population provided that it exists at all in the population at t, i.e. $f_t(m) > 0$, and that it has more than average reproductive success, i.e. $H_t(m) - H_t > 0$.

In what follows we will confine ourselves basically to a static characterization of the stability of the results of evolutionary processes in terms of *e*volutionarily *s*table *s*trategies, ESS and, for that matter, to the closely related concept of *l*imit *e*volutionarily stable *s*trategies, LESS. The concept of an LESS introduced in Selten (1988) is broader than the original notion of an ESS (cf. Maynard Smith and Price 1973) in that

every ESS is a LESS but not vice versa. Though the concept of an ESS is not always compatible with the stability of the replicator dynamics (cf. Weissing, 1991) the two stability concepts will be equivalent in the specific cases that we study.

Referring to the notation introduced in equations (2.1)-(2.3) a mutant or strategy $m \in M$ can be characterized as evolutionarily stable by

(2.4) $H(m, m) \geq H(m^*, m), \forall\, m^* \in M$

and

(2.5) $H(m, m^*) > H(m^*, m^*), \forall\, m^* \in M$ with $H(m, m) = H(m^*, m)$.

According to the first condition (2.4) an evolutionarily stable strategy is adapted optimally to an m-monomorphic environment in so far as the reproductive success of m in a population consisting entirely of m-individuals is at least as high as that of any other strategy $m^* \in M$. Should some other mutant or strategy $m^* \in M$ be as well adapted as m in an m-monomorphic population m^* nevertheless cannot succeed. For, according to (2.5) m will have greater reproductive success than m^* in an m^*-monomorphic as well as in any population composed exclusively of m and m^* individuals.

The systematic relationship between the concept of an ESS and stable situations that are characterized by a distribution f(.) with $\dot{f}_t(m)=0$ for all $m \in M$ is quite obvious. If $\dot{f}_t(m)=0$ holds good for all $m \in M$ then according to equation (2.3) all mutants or strategies m with $f_t(m)>0$ must have the same reproductive success. If several strategies are optimal the process characterized by (2.3) would reduce the frequency of all strategies m^* because according to (2.5) these strategies would fare relatively worse in every population exclusively composed of m and m^* individuals.

2. The game of trust

So called paradoxes of co-operation and rationality have been widely discussed in social theory (cf. for instance anthologies like Barry and Hardin, 1981 or Campbell and Sowden 1985). The prisoner´s dilemma is certainly the most famous example of this genre. Still, taking the prisoner´s dilemma as a paradigm case of problematic social situations is partly misleading (for an effort to characterize the concept of a problematic social situation see Raub and Voss 1986 who expand on the discussion in Harsanyi 1977). The prisoner´s dilemma is constructed as a simultaneous move game and presented in its normal form. This somewhat distracts from the basic fact that social interaction typically is characterized by a sequence of moves in which one player moves first and another second. Even though in game theoretic modeling simultaneity of moves does not depend on the timing but rather on the information about moves some if not the the most important problematic real world interactions seem to differ in the latter respect from standard prisoner´s dilemma interactions (for some related considerations, see Bolle and Ockenfels 1990).

Anybody who has carefully watched bargaining and subsequent exchange between children, the exchange of spies in cold war Berlin or the exchange of hostages knows that second or last mover advantages are of crucial importance. Two children simultaneously grabbing for what has been promised to them in the bargaining process, spies "simultaneously" runnig to the other side of a Berlin bridge or releasing hostages in the last moment in exchange for some means of escape that are also offered in the last moment are examples that all tell the same story: Without commitment power exchange and agreement are precarious. If one side moves first and fulfils its promise, the other one, now in the advantageous position of a second mover, does not have an incentive to comply with the terms of the agreement. The execution of acts is not completely contingent on each other and *therefore* Pareto inferior results tend to emerge.

One should not be mislead by the somewhat exotic character of the foregoing examples. The problematic character of social exchange and the non-self-enforcing character of the agreements involved is most clearly visible in such anarchic or quasi-anarchic situations. But the underlying problem of trust and reward pervades social life throughout. Within any legal relationship those aspects which have not been explicitly dealt with in legally enforceable contractual clauses will tend to give rise to these problems. Social institutions like promise giving or contract enforcement are nothing but extended answers to such problems. The division of labor and more generally reciprocity hinge on solving problems of trust. In particular, individuals can specialize only if they can trust that the specificity of their resources will not be exploited. And, if we follow the maximin of "do ut des" (I give to make you give) how can we trust that the other, after we have done our part, shall do his?

Insisting that the terms exchange and prisoner´s dilemma may generally be used interchangeably Russell Hardin alludes to the fundamental role of trust in social interaction (cf. Hardin 1982, and also Kliemt 1990). This insight suggests that the problem of trust should be at center stage of any strategic analysis of social interaction in general and social co-operation in particular. Its essential aspects may be scrutinized by studying the following *game of trust* (a variant of the stage game of Rosenthal´s, 1981, centipede game):

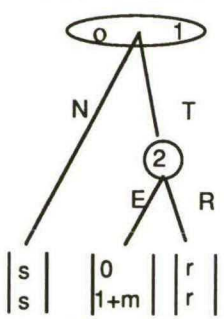

Figure 2.1

The first pay-off at each end node corresponds to the first moving player the second to the second mover. We assume $0<s<r<1$ and $\frac{1}{2} < r$ (where the second requirement is not essential for the subsequent argument but merely excludes cases in which optimal co-operation would require that individuals take turns in exploiting each other). The two moves N, T of player 1 may be interpreted as corresponding to "no trust" and "trust" respectively. Move E refers to "exploitation" of a trusting player 1 while R symbolizes "(fair) reward".

The parameter "m" is purely behavioral. It is interpreted as the effect of a "conscience". It is internal to the actor rather than being influenced by access to resources of the external world. In principle m could be any real number. As choices depend only on the order between the pay-offs it is sufficient, however, to study merely two values - $\underline{m}$, $\bar{m}$ - of the parameter m. The relevant relation is

(2.1) $\underline{m} < r-1 < \bar{m}$.

Consider $m=\bar{m}>r-1$. In this case, which in particular includes m=0, the strategy combination (N, E) is the single subgame perfect equlibrium of the game. Because of r>s playing this way is Pareto-inefficient. On the other hand $\underline{m}<r-1$ implies that the single subgame perfect equilibrium of the game is (T, R). Then both players get a payoff of r. Thus, if the rules of the $\bar{m}$-game are changed such that the $\underline{m}$-game emerges both players shall be better off. The external institutions of promise giving as well as that of contracting may serve the purpose of modifying the preference order of second movers. They may provide means of commitment such that individuals can deliberately choose to modify their future incentives. External punishment of certain acts would do either and so will the emergence of a sufficiently strong conscience "in" the player adopting the role of the second mover. It could also bring about such a change of the rules. Focusing on conscience in this paper this raises the question whether and under which circumstances a (sufficiently strong) conscience can be expected to prevail.

3. The evolution of a conscience

The question may be addressed within the framework of evolutionary game theory. More specifically evolutionary stability of alternative values of the parameter m may be determined for the most simple strategy or mutant space

(3.1) $M=\{\underline{m}, \bar{m}\}$.

The two strategies or mutants $m \in M$ lead to two different *types* of players. We will refer to the $\underline{m}$-type who is endowed with a sufficiently strong conscience as a "player with a conscience" or the "fair player (type)". Correspondingly the $\bar{m}$-type will be called "player without a conscience" or the "unfair player (type)".

Let us generally assume that fair and unfair players indefinitely interact in games of trust. They are matched randomly and in each of the matches they play the game independently of their memory of past or their expectations of future interactions with either the same or other players. Whether they adopt the role of the first or that of the second mover is determined by a random process that is beyond their strategic control. With probability 1/2 each of the two roles may be assigned to them. After role assignment the players know their position in the game.

Under these general assumptions we shall subsequently study the effects of different information conditions. We start with the assumption that players can recognize beforehand and with certainty each other´s type. This leads to what we shall call the "evolutionary game of trust with complete type information".

3.1. The evolutionary game of trust with complete type information

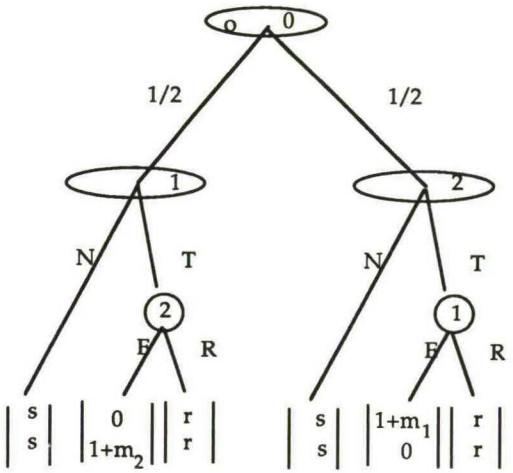

Figure 3.1[1]

Analyzing the game of figure 3.1 from an evolutionary point of view we must determine the relative fitness $H(m_1, m_2)$ for all constellations of parameters (m_1, m_2), $m_1, m_2 \in M=\{\underline{m},\bar{m}\}$. This can be done on the basis of our knowledge of solutions that will emerge in the game of trust for different values of the parameter m (cf. figure 2.1). Given our assumption that players can identify the other player's type with certainty the first moving player knows the second mover's type. The first mover knows whether or not after playing T he can rationally trust that the second mover will retribute by the choice of R. Therefore, after the elimination of dominated strategies, (T, R) is the solution for $m=\underline{m}$ while (N, E) is rational for $m=\bar{m}$. According to (T, R) the pay-off or, biologically speaking, the reproductive success of each player is r while it will be s for each player if (N, E) is the subgame perfect equilibrium.

The results of rational play of the game depend only on the type of the second mover. Each player, regardless of her type, will be in that

[1] Neither in this graph nor subsequently shall we distinguish explicitly between equivalent moves like T, N in different subgames though they are, of course, different moves. Since the distinction is obvious anyway this notational simplification will not lead to confusions.

position with probability 1/2. Therefore we can easily calculate the expected reproductive success of the fair and unfair types of players. The results of the calculations for the four parameter constellations are shown in the following table

m_1 / m_2	$\underline{m}$	$\bar{m}$
$\underline{m}$	r / r	r/2+s/2 / r/2+s/2
$\bar{m}$	r/2+s/2 / r/2+s/2	s / s

Figure 3.2

The entries in the table represent expected reproductive success. Focusing on the row player - entries in the upper left corner of each cell - we may observe that independently of whether $m_2 = \underline{m}$ or $m_2 = \bar{m}$ holds good the fair type of player 1 shall fare better than the unfair type. Because of r>s the reproductive success of player 1 shall be greater under $m_1 = \underline{m}$ than under $m_1 = \bar{m}$. This proves

Theorem 3.1: There is exactly one evolutionarily stable strategy in the evolutionary game of trust with complete type information, namely $\underline{m}$.

Starting with any population that contains both types the $\underline{m}$-type will eventually eliminate the $\bar{m}$-type provided that players, who happen to be in the first mover position, can identify the other player´s type. We may infer therefore from theorem 3.1 that the genetic disposition to develop a sufficiently strong conscience will pay off in the currency of reproductive success if complete type information prevails. Under such ideal information conditions (cf. the related assumption of "trans-

parency" in Gauthier 1986, 174) only the "moral" geno-type can persist in evolution.

The foregoing argument amounts to a rudimentary (potential) evolutionary explanation of how and why a conscience could evolve under ideal information conditions.[2] We now turn to the polar case in which players cannot identify the their partner´s type beforehand. To this case we will refer as the "evolutionary game of trust with incomplete type information".

3.2. The evolutionary game of trust with incomplete type information

That players can identify each other´s type with certainty may be assumed with some plausibility only in small closely knit societies or, more generally, small stable groups (which of course may be part of larger organizations). However under such conditions our general assumption of random matching of players without memory and reputation effects becomes quite implausible. In particular, players should be expected to choose their partners for the game of trust according to their type information whenever possible. If we take into account this the assumption of random matching fits much better with a process of anonymous interaction of a large number of players. Interaction on a large anonymous market would form a typical example. But then it becomes much more plausible to start from the premise that type information is lacking. This quite naturally leads to the assumption that players know the distribution of types in the entire population but do not have any information about the type of their partner.

In our analysis of this case we shall assume more specifically that players do not know the type of their partner if they have to decide between N

---

[2] Even if the play of (T, R) would consume some resources the same result could be reached. Observe that the reproductive success of the $\underline{m}$-type in any case exceeds that of the $\bar{m}$-type by $\frac{r-s}{2}$. Within the clearance left by $\frac{r-s}{2} > 0$ only the critical value of the parameter would change. For additional costs of $\mu>0$ the requirement for a sufficient strength of the conscience $\underline{m}$ would turn into $1+\underline{m}<r-\mu$.

and T. Given that p, $0 \leq p \leq 1$, denotes the share of $\underline{m}$-types in the entire population and 1-p accordingly the share of $\bar{m}$-types, we assume that the belief parameters p and 1-p are common knowledge. Moreover each player knows his own type or his own parameter $m \in M$.

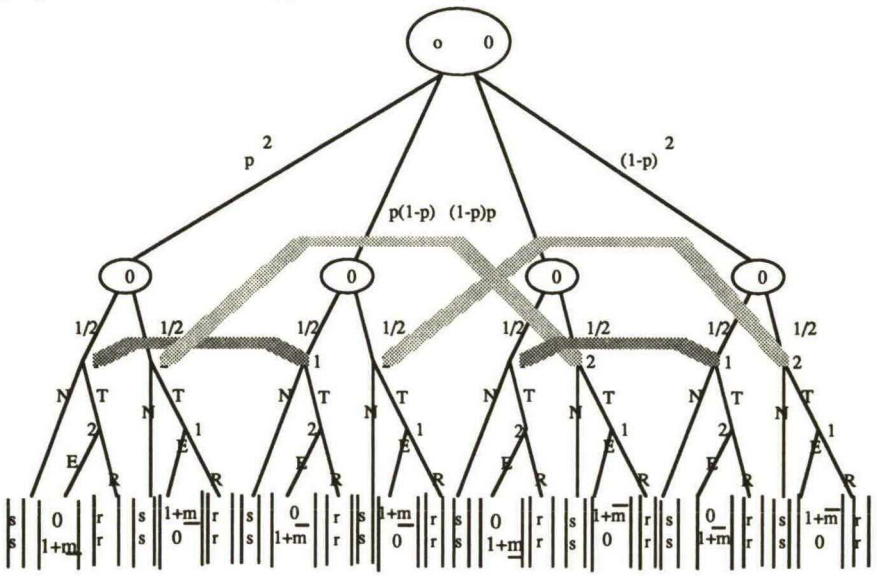

Figure 3.3

Figure 3.3. represents the evolutionary game of trust with incomplete type information. At the start of the game nature as the player numbered 0 makes three random moves. Firstly the type of player 1, secondly the type of player 2 and thirdly the roles of first and second mover are determined randomly. For convenience we have collapsed the random determination of types into one initial random move. With probability 1/2 and independently of their type players will get into the position of first or second mover. Afterwards players make their moves under common knowledge of p but ignorant of each others´ type. The optimal decisions of the second mover do *not* depend on the type of the first mover. Thus the second mover´s information sets become strategically

irrelevant and can therefore be neglected in the graphic presentation of the game.

The information sets of first moving players who are deciding between N and T are symbolized by dotted braids. Deciding whether "to trust or not to trust" the first mover does not know the type of the second mover. For instance, if a fair type player 1 is assigned the role of a first mover he only knows that he is at one of the left decision nodes which are reached with probability p^2 and $p(1-p)$, respectively. Obviously, these probabilities determine the posteriori beliefs of the first mover in an unambigous way: the $\underline{m}$-type is expected with probability p while the $\bar{m}$-type is expected with complementary probability (1-p). For a first moving player 1 who is of the unfair type the corresponding probabilities are $(1-p)p$ and $(1-p)^2$. Analogous considerations apply if player 2 is the first mover while player 1 moves second.

Neglecting the information sets of the second mover as strategically irrelevant we can anticipate her optimal decisions, R for $m=\underline{m}$ and E for $m=\bar{m}$. This leads to the truncated game tree of figure 3.4.

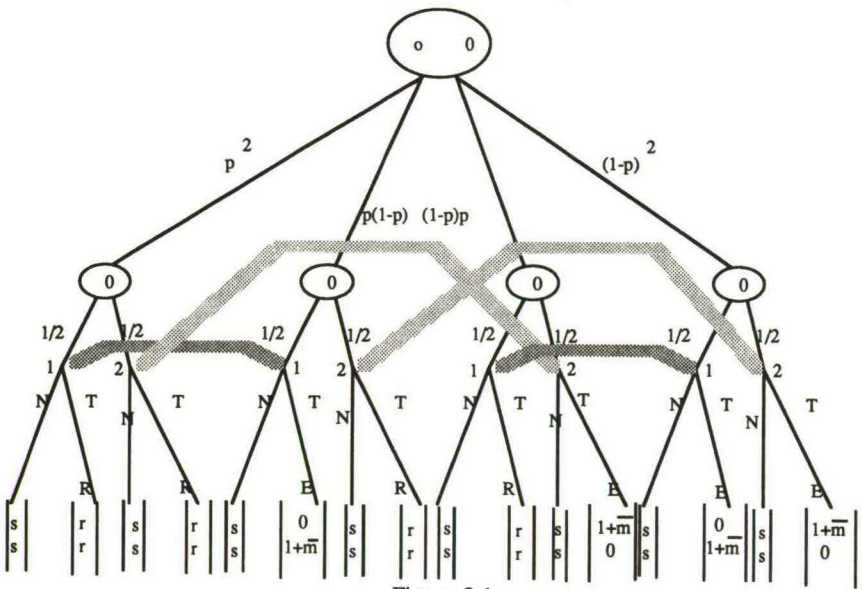

Figure 3.4

Analyzing this tree the expected reproductive success of the mutants $\bar{m}$, $\underline{m} \in M$ can be determined in a fairly simple way. We note first, that individuals who end up as a first mover will independently of their own type choose N if $s>rp + 0(1-p)$ and T if $s<rp + 0(1-p)$. Pursuing this line of argument somewhat further we can generally distinguish two cases: $rp>s$ and $rp<s$.

If $rp>s$ holds good the following two relations determine the expected reproductive success of the two types

$$(3.1)\quad \frac{1}{2}[pr + (1-p)0] + \frac{1}{2}r = \frac{1}{2}(pr+r) \qquad \text{for } m=\underline{m}$$

and $$(3.2)\quad \frac{1}{2}[pr + (1-p)0] + \frac{1}{2}1 = \frac{1}{2}(pr+1) \qquad \text{for } m=\bar{m}.$$

Consider, for instance, the left side of relation (3.1). With probability 1/2 a player of type $\underline{m}$ will be assigned the role of a first mover. Because of $rp>s$, he will choose T then. With probability p the second mover will be of type $\underline{m}$, too. This type will fairly reward trust by move R. Consequently the first mover shall receive r with probability p. With probability (1-p) the second mover will be of the unfair type. Trust shall be met with exploitation, E, and therefore will yield 0 for the first mover. With probability 1/2 the $\underline{m}$-type will play the game as a second mover. As the first mover independently of his own type will choose T if $rp>s$ prevails the $\underline{m}$-type who prefers R to E can expect r. A completely parallel argument can be used in interpreting relation (3.2). This concludes the analysis of case $r>ps$.

If $rp<s$ holds good the first mover shall independently of her type choose N. There will be no second move and the pay-offs for both types of players shall be s in this case.

Since $r<1$ the value determined in (3.2) is larger than the one determined in (3.1). Thus the expected reproductive success of $\bar{m}$ in an $\underline{m}$-monomorphic population is larger than that of $\underline{m}$. An $\underline{m}$-monomorphic

population is not evolutionarily stable and neither is any population with $p > \frac{s}{r}$. This discussion is summed up by

Theorem 3.2: In the evolutionary game of trust with incomplete type information no population composition with a share $p > \frac{s}{r}$ of fair $\underline{m}$-types can be evolutionarily stable.

Theorem 3.2 states that in any population composed exclusively of fair types the unfair types would have a higher reproductive success once they enter the population. Still, once the percentage p of individuals of type $\underline{m}$ shrinks to some value lower than $\frac{s}{r}$ there would be no evolutionary pressure in favor of unfair types anymore. First movers would choose N all the time. The fact that fair types prefer R over E would become irrelevant. Beyond the threshold $\frac{s}{r}$ the disposition to behave fairly would not be weeded out.

However, the absence of evolutionary pressure presupposes that players can behave in ways that perfectly comply with the precepts of rational choice. In a complex world in which phenotypes must pursue complex behavioral programs it may be more adequate to assume that individuals regardless of their faculty to choose rationally (a biological endowment as well) will make slight mistakes or non-rational choices once in a while. Under this assumption we can strengthen the concept of an ESS to that of a *l*imit *e*volutionarily *s*table *s*trategy (LESS; cf. Selten 1988) which captures the notion of slight mistakes by the mathematical idea of probabilistic perturbations.

According to these perturbations even those moves that rationally would be performed with probability 0 shall be performed with some sufficiently small probability $\varepsilon > 0$. Evolutionary stability of such perturbed games of trust can be studied and the limit $\varepsilon \to 0$ can be taken. Again it is helpful to distinguish between the two cases $rp > s$ and $rp < s$.

In case rp>s the relation (3.2)>(3.1) will hold good in perturbed games. The difference in reproductive success will not be compensated by small perturbations. Therefore reproductive success will be type dependent in (slightly) perturbed games. Minimum probabilities for E and R would merely narrow the scope for the population composition p.

In case rp<s expectations were not type dependent in the unperturbed game. However in the perturbed game expectations become type dependent.

Let ε be a small, but positive minimum probability for T as a first mover's choice. We get

$$(3.3)\ \frac{1}{2}[(1-\varepsilon)\,s + \varepsilon\, rp] + \frac{1}{2}[(1-\varepsilon)\,s + r\varepsilon] \text{ for } m=\underline{m}$$

and

$$(3.4)\ \frac{1}{2}[(1-\varepsilon)\,s + \varepsilon\, rp] + \frac{1}{2}[(1-\varepsilon)\,s + 1\varepsilon] \text{ for } m=\bar{m}.$$

If they are assigned to the role of a first mover both types have the same expectations. For, in that role, both independently of their own type shall perform move N with the perturbed probability $(1-\varepsilon)$ and move T with minimum probability ε. The terms in the first brackets of (3.3) and of (3.4), respectively, coincide. Consider now the last terms of the second brackets of (3.3) and of (3.4) respectively. According to these terms which refer to the pay-off in the role of a second mover the $\bar{m}$-type fares better receiving 1 with probability ε while the fair $\underline{m}$-type shall get merely r<1 with probability ε. This implies that the comparative advantage of the unfair $\bar{m}$-type carries over from the realm rp>s into rp<s. Thus, if move T cannot be excluded with certainty only populations with a percentage p=0 of fair $\underline{m}$-types shall be evolutionarily stable.

Moreover, a positive minimum probability of move N has no effect if rp<s prevails. In that case N should be rationally performed with probability 1 anyway. We can therefore note:

Theorem 3.3: In the evolutionary game of trust with incomplete type information only an $\bar{m}$-monomorphic population with

p=0 can be evolutionarily stable according to the LESS concept.

Comparing theorems 3.1 and 3.3 the importance of type information becomes obvious. While developing a conscience is evolutionarily stable if players can identify their co-players' types before they play a game of trust this is no longer true if players merely know the distribution of types in the entire population. If types cannot identify each other, a first mover cannot expect that trust shall be met with fair retribution by the second mover. The disposition to develop a conscience that is sufficiently strong to motivate fair behavior shall not succeed in evolution. Only the inclination to behave unfairly in interactions that have the incentive structure of the game of trust shall be evolutionarily stable.

The foregoing arguments were related to pure, ideal or extreme cases. As such they give us some handles on what might have been crucial factors of the evolutionary process in which our own species emerged. Nevertheless it seems quite unlikely that the ideal conditions of the two polar cases could have prevailed during human evolution. It seems much more plausible that these conditions as a matter of fact can be characterized by "intermediate" informational assumptions that locate the interaction between the two extreme cases studied so far. To an analysis of such an intermediate case we shall turn now.

3.3. The game of trust with some type information

In the real world information is a valuable commodity. Situations that have the structure of the game of trust illustrate this general observation in a particularly interesting way. For, in such situations the availability of information about the trustworthiness of second movers is crucial. Since r>s, first as well as second movers should be willing to invest in costly information technologies. First movers independently of their own

type would like to detect the type of the second mover while trustworthy second movers would like to signal their own type. As the comparison of the two polar cases analyzed before shows, it would be in the common interest of the first and the trustworthy second mover if the type of the second mover could be made common knowledge by some signal.

However, a second mover without a conscience has an incentive to engage in mimicry and this in turn puts a threat on the reliability of signaling and detection technologies. Since we cannot dwell on the biological aspects of this and related issues here we shall simply assume that a reliable detection technology is available at a cost. Individuals can make a strategic decision whether or not they shall invest in that costly detection technology before they play the game of trust.

These investment decisions determine the information conditions that prevail in the game of trust. If both invest the type of both players shall be common knowledge and the game of trust with complete type information shall be played. If none invests merely the distribution of types in the population shall be common knowledge and the game of trust shall be played without any specific type information. If exactly one individual invests which of the games is played depends. If the investing individual ends up as a first mover the emerging game of trust shall have basically the same strategic properties as the game with complete type information. If the investor plays as a second mover there will be no return on the information investment and the emerging game has basically the same strategic properties as the game of trust without type information.

3.3.1. The game model

Observe first that independently of information conditions, the second mover´s final decision between E and R shall depend only on her own type. Irrespective of what the last mover knows about her own type as

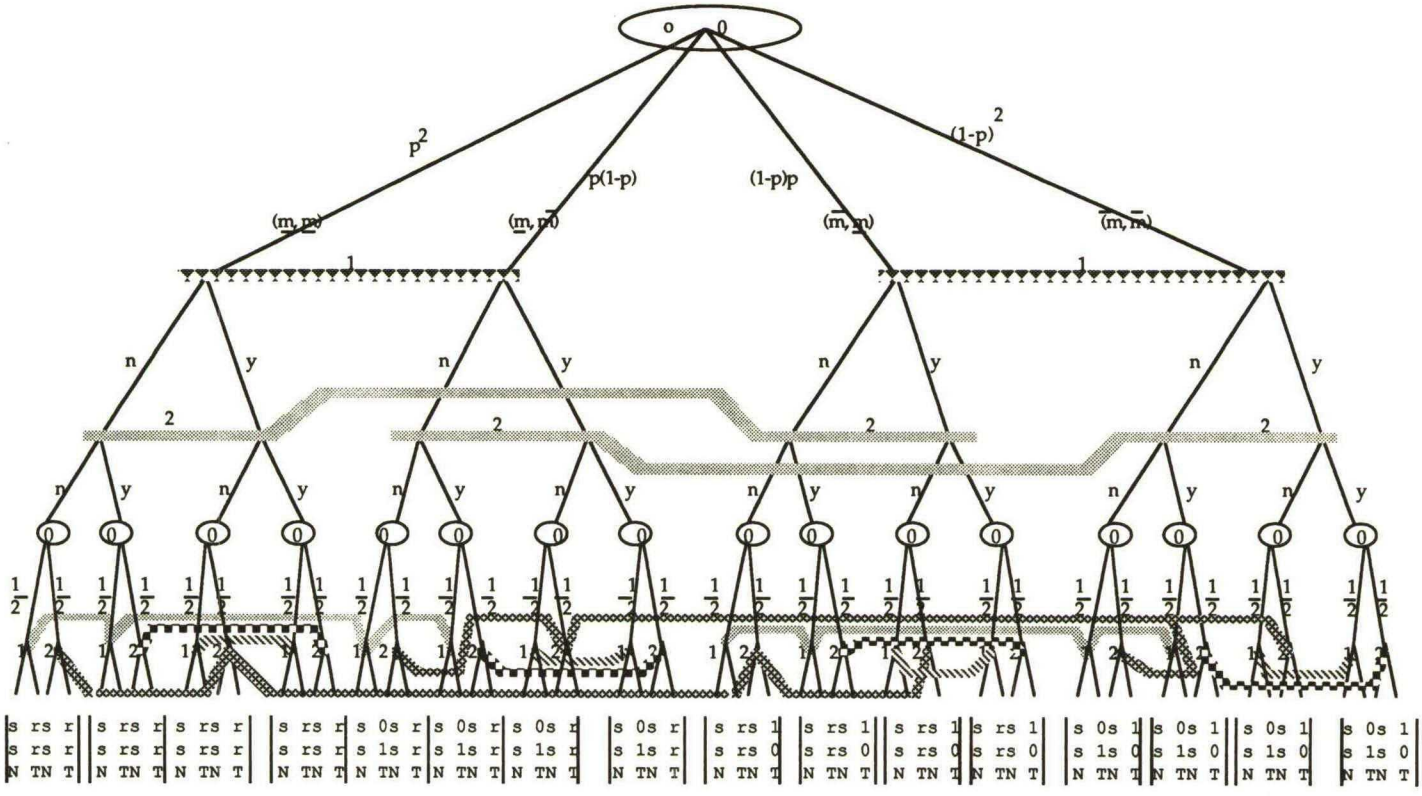

Figure 3.5: In this truncated tree it is assumed that the second mover chooses optimally according to his type after "T". After a player has chosen y the cost C must still be subtracted from all payoffs of that player at the corresponding endnodes. The information sets are indicated by braids. The last line, showing whether "T" or "N" was chosen by the first mover, as well as the vertical lines separating "packs" of four payoff vectors each were included for the convenience of the reader.

well as about the first mover´s type and moves a last mover of type $\underline{m}$ shall rationally choose R while her $\bar{m}$-type prefers E. Thus without changing essential strategic elements of the game, we can substitute the final decisions of the second mover by the payoff vectors resulting from rational play and we can present the game by a truncated variant of its tree (see figure 3.5).

figure 3.5 about here

Again the game starts with a fictitious initial chance move. In this move the chance player 0 selects both players´ types. Independently of each other with probability p the $\underline{m}$-type and with complementary probability (1-p) the $\bar{m}$-type are chosen (where p is the proportion of $\underline{m}$-types in the population). Then players must decide whether they shall invest, y, or not invest, n, into the detection technology. Both can invest at the same cost "C". They must make that decision before the roles of a first or of a second mover are assigned to them. When they make their decisions they are ignorant not only of the other player´s type but also of the other´s investment decision. For every constellation of m-types and players´ investments into detection technologies players are then assigned to their roles as first and second mover with probability 1/2 by an unbiased chance move. Finally under the given constellation a variant of the game of trust is played resulting in payoffs modified by the constant C. (The payoffs of the truncated tree of figure 3.5 must still be modified by subtracting the investment costs C of any player who chose y. We left this out because otherwise the game tree might have become virtually unreadable.)

According to the information sets of figure 3.5 the players´ decisions between y and n remain private information. Since in the basic game of trust the information technology can be useful only for the player who moves first this is not crucial. The first mover´s optimal decision does not depend on the information status of the second mover while the

second mover would not care anyway. Thus, it would not change the strategic character of the game in any essential way if investment decisions could be observed or were anounced beforehand. More restrictive is the assumption that the information technology is perfectly reliable. After investment the investing player knows the other player's type. The costs may be too high to make the investment worthwhile. However, if the costs of the technology are borne it works perfectly reliable. For instance, if both have invested in detection technology the game under perfect type information shall emerge (though all payoffs in the basic game of trust that is reached after these investment decisions are diminished by a constant C).

3.3.2. The game of trust with incomplete type information if a perfect detection technology is available at a cost

Taking into account what has been said before about the strategic situation of second movers we may focus first on the decision of an individual who, with probability 1/2, ends up as a first mover. Assume that the first mover had chosen y. He then gets to know the m-type of his co-player. Therefore, if the co-player is of type $\underline{m}$ he will choose T whereas facing an $\bar{m}$-type opponent he will play N. Had he chosen n instead of y the first player's decisions would depend on his a priori beliefs in the following way: He would prefer T if rp>s and N if rp<s.

The payoff implications of n depend on whether rp>s or rp<s holds good. Therefore we will distinguish between these two cases when analyzing the optimality of investment decisions for players who do not yet know whether with probability 1/2 they shall play as first movers or with equal probability shall be in the position of a second mover.

Case rp>s: In this case a player who chooses n and then -- with probability 1/2 -- must move first, on behalf of his a priori beliefs shall opt for T. Clearly, if -- with probability p -- the opponent is of the $\underline{m}$-

type, the choice of n (except for C) shall yield the same payoff expectation as that of y. For, against an $\underline{m}$-type opponent, T would be optimal after y as well. If, on the other hand, the opponent -- with probability 1-p -- is of type $\bar{m}$ choosing y will yield s whereas 0 accrues after the choice of n. Choosing y enhances the payoff expectation only when playing the basic game of trust as a first mover. Since this applies with probability 1/2 while the detection technology is useless if the player ends up as a second mover the expected advantage of y over n is $\frac{1}{2}$(1-p)s which must be compared with the cost C. We note

Lemma 3.1: For rp>s the choice of y is optimal for $(1\text{-}p)\frac{s}{2} > C$ whereas n is better if $(1\text{-}p)\frac{s}{2} < C$.

Case rp<s: In this case a player who chooses n and then -- with probability 1/2 -- must move first shall opt for N on behalf of his a priori beliefs. If he encounters an $\bar{m}$-type opponent -- which happens with probability (1-p) -- the payoff expectations after y and n (disregarding C) shall be equal. Then knowledge of the other player's type shall not alter the first mover's choice. If, on the other hand, the opponent is an $\underline{m}$-type -- an event pending with probability p -- the return on a positive investment decision y (at cost C) shall be r whereas the payoff accruing to the choice of n shall merely amount to s. We note

Lemma 3.2: For rp<s the choice of y is optimal for $(r\text{-}s)\frac{p}{2} > C$ whereas n is better if $(r\text{-}s)\frac{p}{2} < C$.

Lemmas 1 and 2 jointly with our previous results determine a solution of the detection model of the truncated game tree of figure 3.5 for all generic parameter constellations that exclude indifference along the solution path. This may be summarized in

Theorem 3.4: Except for degenerate cases in which at least one player is indifferent between alternative moves along the solution path the detection model characterized by the truncated game tree of figure 3.5 has a unique solution according to which

(i) in the basic game of trust a second mover who is of type $\tilde{m}$ chooses E while in that role the $\underline{m}$-type prefers R,

(ii) in the basic game of trust a first mover who has made the positive investment decision y knows the second player's type and therefore chooses T if her co-player is of type $\underline{m}$ and N otherwise,

(iii) in the basic game of trust a first mover who has made the negative investment decision n chooses T if $rp>s$ while preferring N if $rp<s$,

(iv) the optimal investment decision of both players is y if

$$rp>s \text{ and } (1-p)\frac{s}{2} > C$$

or

$$rp<s \text{ and } (r-s)\frac{p}{2} > C$$

whereas the optimal investment decision of both players is n if

$$rp>s \text{ and } (1-p)\frac{s}{2} < C$$

or

$$rp<s \text{ and } (r-s)\frac{p}{2} < C.$$

Theorem 3.4 deals merely with non-degenerate cases. Degenerate cases emerge whenever one of the players is indifferent between two moves. Given optimal play this may happen only for highly special parameter

constellations, namely rp=s, $p=\frac{s-2C}{s}$, and $p=\frac{2C}{(r-s)}$. It shall become obvious in the next section that except for $p=\frac{s-2C}{s}$ these values of p refer merely to transitory states of evolutionary dynamics and thus are irrelevant for the determination of the evolutionary stable population composition.

3.3.3. The evolutionary game of trust with incomplete type information if a perfect detection technology is available at a cost

Before turning to the issue of evolutionary stability itself it seems helpful to take a somewhat closer look at the relationships between the cost C of the detection technology and the solution behavior. Because of $p\in[0, 1]$ and r>s>0 the expected gains from investment $(1-p)\frac{s}{2}$, in case rp>s, and $(r-s)\frac{p}{2}$, in case rp<s, are both smaller than $\frac{r}{2}$. Thus for $C\geq\frac{r}{2}$ the negative investment decision n is optimal in any case.

Taking into account theorem 3.4 (iv) a more stringent condition can be derived. For $C\in(0, \frac{r}{2})$ the highest possible gain from detection is $(1-\frac{s}{r})\frac{s}{2}$ $=(r-s)\frac{s}{r}\frac{1}{2}$. It is achieved for $p=\frac{s}{r}$; i. e. when a player who is assigned the role of a first mover in the basic game of trust is indifferent between T and N. From this we can infer that n is optimal if $C\geq\frac{s}{2}(1-\frac{s}{r})$. Since 1>r>s>0 implies $\frac{r}{2}>\frac{s}{2}(1-\frac{s}{r})$ this condition is indeed more stringent than the one derived before.

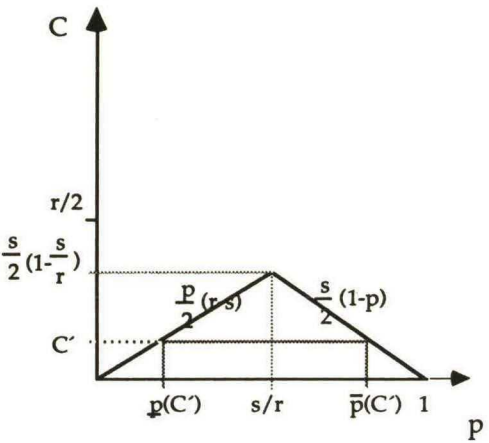

Figure 3.6

Within the range $\frac{s}{2}(1-\frac{s}{r})>C>0$ the result depends on the population composition p as illustrated in figure 3.6 for the value C´. For all p with $\underline{p}(C')<p<\bar{p}(C')$ expected gains from a positive investment decision y exceed their costs C´ whereas for any p outside that range the opposite holds. Thus investing in detection technology is worthwhile if for C=C´ the population composition complies with $\underline{p}(C')<p<\bar{p}(C')$ where $\underline{p}(C')$ and $\bar{p}(C')$ are determined by

$$\underline{p}(C')=\frac{2C'}{r-s} \text{ and } \bar{p}(C')=\frac{s-2C'}{s}.$$

More generally we can state for all cost parameters $C\in(0, \frac{s}{2}(1-\frac{s}{r}))$ that the positive investment decision y can be rational only if the population compostion p complies with

$$\underline{p}(C)=\frac{2C}{r-s}<p<\bar{p}(C)=\frac{s-2C}{s}.$$

For all $C\in(0, \frac{s}{2}(1-\frac{s}{r}))$ we can finally infer -- cf. also figure 3.6 --

$$0<\underline{p}(C)<\bar{p}(C)<1.$$

Keeping in mind these preliminary observations about the relationships between the costs of the detection technology, the population composition and optimal investment behavior we can now approach the problem of evolutionary stability. As before we shall only consider the mutant space $M=\{\underline{m},\bar{m}\}$ with $\underline{m}<r-1<\bar{m}$ where we can and shall rely on our previous results whenever suitable.

According to part (iv) of theorem 3.4 the optimal investment decision does not depend at all on the investing player's type. This is intuitively plausible. On the one hand, the faculty to detect the other player's type is for first movers of both types of the same value while their optimal choice does not depend on their own type. On the other hand, though the optimal decisions of second movers are type dependent the detection technology is useless for them. Therefore the probability of choosing y is determined by considerations that are identical for both types. We shall refer to this type independent probability of choosing y by $x\in[0, 1]$.

Consider first of all prohibitively high values $C > \frac{s}{2}(1-\frac{s}{r})$. Then $x=0$ will hold good. Knowledge of m-types is strictly private and a priori beliefs are determined by the true population composition p. Players shall only know their own m-type and the distribution of m-types in the entire population. Therefore theorem 3.3 directly applies. The population share p of $\underline{m}$-types converges to $p=0$. The evolutionarily stable population composition does not contain trustworthy individuals. Conscience cannot survive.

Lemma 3.3: In the evolutionary game of trust with perfect but costly detection technology only an $\bar{m}$-monomorphic population with $p=0$ can be evolutionarily stable according to the LESS concept if $C > \frac{s}{2}(1-\frac{s}{r})$. In such a population no one invests in detection technology.

For cost parameters $C\in(0, \frac{s}{2}(1-\frac{s}{r}))$ considerations become more complicated. Whether or not the costly detection technology will be used and

whether or not the evolutionarily stable population composition will contain $\underline{m}$-types then depends: Whereas investing in detection technology is optimal if $p \in (\underline{p}(C), \bar{p}(C))$, choosing n is optimal for $p \notin (\underline{p}(C), \bar{p}(C))$.

Since $0<\underline{p}(C)<\bar{p}(C)<1$ we know that monomorphic population compositions fall outside that range. We can therefore note two simple lemmas.

Lemma 3.4: In the evolutionary game of trust with perfect but costly detection technology an $\bar{m}$-monomorphic population with p=0 will always be evolutionarily stable according to the LESS concept if C>0. In such a population no one invests in detection technology.

Proof: In case of p=0 the gain of using the detection technology shall be zero (cf. again figure 3.6). Because of C>0 we therefore can infer that x=0 and the situation with no investment in detection technology prevails. Therefore lemma 3.4 follows from theorem 3.3 ∴.

Thus, due to C>0 and (r-s) $p\frac{1}{2} \rightarrow 0$ for $p \rightarrow 0$ the $\bar{m}$-monomorphic population with p=0 shall always be evolutionarily stable according to the LESS concept. However, the other possible monomorphic population, the $\underline{m}$-monomorphic one with p=1 is never evolutionarily stable.

Lemma 3.5: In the evolutionary game of trust with perfect but costly detection technology an $\underline{m}$-monomorphic population with p=1 can never be evolutionarily stable according to the LESS concept if C>0.

Proof: If p=1 nothing can be gained by using the detection technology. Because of C>0 we therefore can infer that x=0. Thus lemma 3.5 again follows from theorem 3.3 ∴.

Turning to other population compositions p observe first that $0<C<\frac{s}{2}(1-\frac{s}{r})$ and $p \notin [\underline{p}(C), \bar{p}(C)]$ implies x=0 and thus a decrease in p. This shows

that no population composition p can be evolutionarily stable under these premises. It remains to be investigated whether population compositions $p \in [\underline{p}(C), \bar{p}(C)]$ can be evolutionarily stable if the detection technology is costly but not prohibitively so.

The interval $[\underline{p}(C), \bar{p}(C)]$ is non-empty for $C \in (0, \frac{s}{2}(1-\frac{s}{r}))$. From the argument illustrated in figure 3.6 one can infer that x=1 if $p \in (\underline{p}(C), \bar{p}(C))$. In that case the m-types shall be common knowledge when the basic game of trust is played and theorem 3.1 implies that p approaches 1. Again refering to figure 3.6 it is obvious that the increase of p has to stop at $p=\bar{p}(C)$. For, beyond $\bar{p}(C)$ nobody invests in detection technology anymore. For $p>\bar{p}(C)$ type information will be private when the basic game of trust is played and the reproductive success of $\bar{m}$-types shall exceed that of $\underline{m}$- types. Switching from x=1 for $p \in (\underline{p}(C), \bar{p}(C))$ to x=0 for $p>\bar{p}(C)$ stabilizes the population composition at $p=\bar{p}(C)$. This proves

Lemma 3.6: For $C \in (0, \frac{s}{2}(1-\frac{s}{r}))$ a population composition p with a share

$$p=\bar{p}(C)=\frac{s-2C}{s}$$

of $\underline{m}$-individuals is evolutionarily stable.

The other boundary $\underline{p}(C)$ of the interval $[\underline{p}(C), \bar{p}(C)]$ does not satisfy the conditions for this kind of evolutionary stability. Since x=0 for $p<\underline{p}(C)$ implies that p decreases and x=1 for $\underline{p}(C)<p<\bar{p}(C)$ induces an increase of p -- where $0<\underline{p}(C)<\bar{p}(C)<1$ due to $0<C<\frac{s}{2}(1-\frac{s}{r})$ and $0<s<r$ -- the population composition $p=\underline{p}(C)=\frac{2C}{(r-s)}$ is highly unstable. Any slight disturbance of p will lead away from $p=\underline{p}(C)$.

Our discussion of the evolutionary game of trust with incomplete type information if a perfect detection technology is available at a cost C may be summarized now in

Theorem 3.5: The evolutionary stability of population compositions characterized by the share p of $\underline{m}$-individuals depends on the costs C of the detection technology as follows:

(i) for $C \geq \frac{s}{2}(1-\frac{s}{r})$ the only evolutionarily stable population is the $\bar{m}$-monomorphic one characterized by p=0;

(ii) for $0<C<\frac{s}{2}(1-\frac{s}{r})$ there are two evolutionarily stable populations characterized either by p=0 or $p=\bar{p}(C)=\frac{s-2C}{s}$;

(iii) for C=0 the only evolutionarily stable population is the $\underline{m}$-monomorphic one characterized by p=1.

Proof: For $C>\frac{s}{2}(1-\frac{s}{r})$ part (i) merely restates lemma 3.3. If $C=\frac{s}{2}(1-\frac{s}{r})$ the interval $[\underline{p}(C), \bar{p}(C)]$, though non-empty still, contains only one point, namely $p=\frac{s}{r}$. Then y is at most as good as n but definitely not better than n. Assume, nevertheless, x=1 for $p=\frac{s}{r}$. Since x=0 for all $p\neq\frac{s}{r}$ one can hardly imagine a dynamic process in which the parameter p as a function of time t with p=p(t) stabilizes at $p=\frac{s}{r}$. Note in particular that x=0 is as plausible as x=1 for $p=\frac{s}{r}$. Therefore one should expect that only p=0 is evolutionarily stable for $C=\frac{s}{2}(1-\frac{s}{r})$. (ii) follows from lemmas 3.4 and 3.6. (iii) restates theorem 3.1.∴

4. Discussion

We are not living in a morally perfect world in which trust finds its fair reward whenever shown and exploitative behavior is unknown. Quite to the contrary we know that the human faculty to make forward looking

opportunistic choices renders precarious or even impossible many forms of mutually advantageous co-operation. The general line of the preceding argument may give us some first clues for a somewhat deeper understanding of why this is so. At the same time it draws attention to circumstances under which a conscience and retributive emotions may serve in evolutionarily stable ways as a potential remedy for certain problems of co-operation in a "competitive world".

Admittedly these as well as the following observations have a distinctly speculative flavor. On the other hand they are not "mere" speculations but speculations guided by models. Besides running experiments and testing models against statistical evidence such kinds of guided speculation are definitely a legitimate part of social science research or more traditionally speaking of "moral science". So let us finally on the basis of our models engage in some partly speculative moral science discussion.

Immanuel Kant in a strange combination of Latin and Greek terms spoke of the "homo noumenon" and the "homo phänomenon" (cf. 1798/1991, § 49, E., 158). He argued that moral duties like those of promise keeping must hold without exception. They have a legitimate claim on us "whatever the consequences". If one can swallow this idea of a noumenal world (noumenale Welt) or a world of reason in which action stems from reason itself this may not be without some plausibility. On the other hand we more pedestrian fellows are not living in the world of reason. And as "worldly philosophers" (cf. Heilbroner´s well known book title 1953/83) we have a natural inclination to believe that men in general are pursuing their worldly aims rather than following the precepts of Kantian reason. Phenomena like emotions and passions form the springs of actions. They determine human preferences which in turn determine the course of human affairs.

Still, even such a world could be morally perfect if individuals would all the time be guided by the "right" moral emotions. In particular the individuals´ conscience could conceivably induce compliance with moral norms of fairness in each instance. Because of this preference

modification the individually rational and the moral choice could coincide throughout.

Insights like lemma 3.5 may count as evidence against such dreams of a pre-stabilized moral harmony. Though forbearing from exploitation may comply with the precepts of Kantian reason the emotional disposition to act that way shall not be evolutionarily stable if information is costly. On the other hand, theorem 3.5 provides some more realistic hope that the world though definitely not inhabited exclusively by saints may neither be the playground of a population of devils. If the cost C of the detection technology is not prohibitively high there exists a stable mixed population composition with a share $p=\bar{p}(C)$ of fair $\underline{m}$-individuals which is positive but smaller than 1.

Whereas part (i) and (iii) of theorem 3.5 are merely restating previous results in the more general framework of the detection model, part (ii) adds a new insight since it points out that under certain values of the parameters different stable population compositions are viable. To us this part of the theorem seems rather convincing. Restating Robert Frank´s informal modern version of a "theory of moral sentiments" (cf. 1988, chap. 3, and of course Smith 1759/1966) in more precise terms of evolutionary game theory it predicts that fair individuals will be driven out of a population whenever the share of $\underline{m}$-type individuals falls below a certain threshold $p=\underline{p}(C)$. If the population share p of $\underline{m}$-types exceeds $\underline{p}(C)$ the fair types can survive but cannot eliminate unfair types beyond $p=\bar{p}(C)<1$.

Moreover, a somewhat closer look at the expressions defining the relevant thresholds of p leads to intuitively plausible results. Starting with the expression $p=\underline{p}(C)=\frac{2C}{(r-s)}$ we immediately observe that the requirement necessary for the evolutionarily stable survival of $\underline{m}$-types becomes more severe if $(r-s)\rightarrow 0$. This makes good sense since, then, bearing the risk of being exploited becomes less worthwhile. Quite analogous considerations apply as far as the growth of C in this term is concerned.

The other threshold $p=\bar{p}(C)=\frac{s-2C}{s}$ is more interesting since it refers to the evolutionarily stable composition of the mixed population containing both types. The $p=\bar{p}(C)$ population seems to be quite in line with our experience: There are always some people whom we can trust and some who cannot be trusted; some shall exploit us if we fail to recognize them while others will be fair even though we may also fail to recognize them beforehand.

Theorem 3.5 (ii) justifies at least some optimism that some trustful co-operation can survive once fair types have somehow gained a sufficient share $\underline{p}(C)$ in a population. This seems to presuppose information conditions under which information costs are sufficiently close to zero. From this point of view it is certainly no surprise that in their original adaptation humans lived in small closely knit societies in which information on each others´ behavior is available virtually costless most of the time. That privacy is alien to primitive people may be counted as supporting evidence (cf. also Posner 1981, chap. 6). It may also be observed that virtually all ordered large group interaction is ordered by a skeleton of small groups in which individuals know each other quite well and can eliminate unfair types from the group once in a while (for example the managers of a large company organizing their and other company members´ business in hierarchies of small groups may exclude unfair types).

Besides reducing information costs small group organization has several other aspects. Still the aspect of reducing these costs is important. In the real world we often take some effort to reduce C by means of small group organization. It fits in nicely that in our model

$$\bar{p}(C)=\frac{s-2C}{s}$$

approaches 1 if $C\rightarrow 0$.

In view of lemma 3.5 theorem 3.1 seems to be a highly special result. If our other assumptions apply an $\underline{m}$-monomorphic population can be evolutionarily stable only if a perfect detection technology is available

for free and thus the condition $(1-p)\frac{s}{2} \geq C$ is trivially satisfied. On the other hand, looking at the next figure 3.7 we may also regard theorem 3.1 as a limiting case of the $\bar{p}(C)$-composition for $C \to 0$.

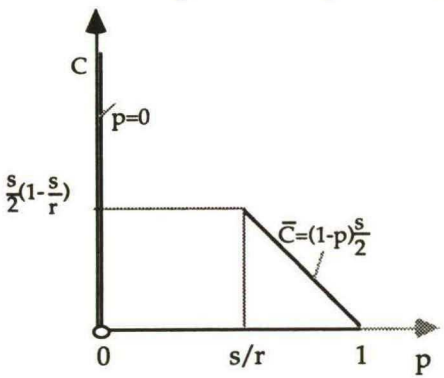

Figure 3.7

The graph summarizes the basic insights of theorem 3.5 in the p,C-plane. Solving $\bar{p}(C)=\frac{s-2C}{s}$ for C we get for each composition p the highest investment $\bar{C} = (1-p)\frac{s}{2}$ in detection technology rational individuals would bear. Obviously with increasing p costs must be lower or with decreasing costs a higher share p of $\underline{m}$-types can be stabilized. The limiting case C=0 is special in the interval beyond p=s/r only insofar as merely one population can be evolutionarily stable then, namely p=1.

For each value of C our analysis determines the evolutionarily stable states, namely either p=0 or p=0 and $p=\bar{p}(C)$, for $C<\frac{s}{2}(1-\frac{s}{r})$. For any initial value $p_0 \in [0, 1]$ it also provides a clue to which of the two composition parameters p will converge. As indicated by the arrows in figure 3.8 the share p of fair types increases inside the triangle formed by the $\underline{p}(C)$-line, the $\bar{p}(C)$-line and the p-axis whereas p decreases outside this realm.

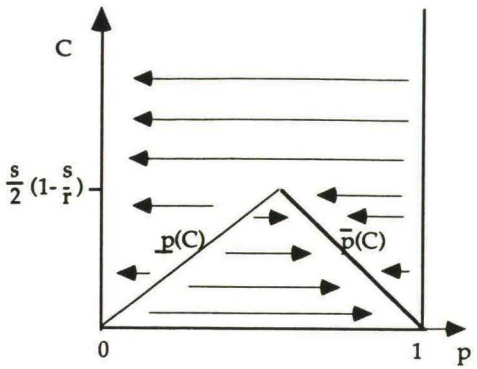

Figure 3.8

The region $\{ (p, C): \underline{p}(C)<p\leq 1, 0\leq C<\frac{s}{2}(1-\frac{s}{r}) \}$

forms the attraction set of $p=\bar{p}(C)=\frac{s-2C}{s}$ for $0\leq C< \frac{s}{2}(1-\frac{s}{r})$ while the region that is complementary to the triangle in figure 3.8 forms the source of p=0 (disregarding degenerate cases with a starting point $p_o=\underline{p}(C)$ and $0<C<\frac{s}{2}(1-\frac{s}{r})$).

According to figure 3.8 the starting point of the dynamic process is essential. If $0<C<\frac{s}{2}(1-\frac{s}{r})$ it depends on the starting point $p_o\in[0, 1]$ whether the population composition will converge to p=0 or to $p=\bar{p}(C)$. Both is possible. If it so happens that we have a "good start" there will be a more or less happy ending if not so not. Since we are talking about biological evolution it does not seem that we can do much about it.

On the other hand, biological evolution shall be influenced by social or cultural evolution in what is basically a co-evolutionary process (cf. on this Lumsden and Wilson 1981, Cavalli-Sforza and Feldman 1981). And, though we cannot directly change our human nature we can do something about social conditions and constraints under which it

operates (cf. on this already Hume 1751/1978, in particular book III, part II, sec. vii). In particular we can influence whether we are living and interacting most of the time in small closely knit groups like the family and stable (typically culturally homogeneous) social networks in which m-types may be expected to be common knowledge or whether we interact more frequently in large anonymous groups. It seems that while the former conditions will stabilize the share p of fair m-types in a population the latter will work in the opposite direction. We therefore should be aware that even in the modern world it may be necessary to stick to some aspects of the social organization that prevailed during the our original hunter gatherer adaption for otherwise evolution might in the long run drive out those moral emotions that make social organization viable in the first place.

A morally perfect world in which people trust each other and in which their trust in all likelihood is met by spontaneous acts of fair reward can be approached as closely as we wish to if we can manage to reduce C. It should also be observed, however, that reducing C may come at an extra cost. For, obviously, whatever may be done to reduce C may not be in line with the fundamental ideals of privacy in a liberal society. Moreover, according to the common wisdom of economists "there is no such thing like a free lunch" and neither is there a free detection technology. Therefore we must in general assume $C>0$. But, then, $p=1$ implies $x=0$. Thus a mutant $\bar{m}$ entering an m-monomorphic population will be successful. In the role of second mover the mutant achieves a reproductive success of

$$\frac{1}{2}r + \frac{1}{2}1 = \frac{1}{2} + \frac{1}{2}r$$

whereas the reproductive success of m-types related to that role is only $r/2$ since trustful and trustworthy first movers in the game of trust fall prey to second mover exploitation but never exploit first movers.

An m-monomorphic population is not evolutionarily stable unless information about the other's type is available at no cost. Since the latter is unrealistic we can end with a kind of inversion of Abraham Lincolns

famous exclamation that one can fool all people sometimes and some people all the time but hardly all people all the time: One can trust all people sometimes, and some people all the time but it would be foolish to trust all people all the time.

References

Alexander, R. 1978: Darwinism and Human Affairs. Seattle and London: University of Washington Press.

Barry, B. and Hardin, R. (ed.) 1982: Rational Man and Irrational Society? Beverly Hills et al.: Sage.

Bolle, F. and Ockenfels, P. 1990: Prisoner´s Dilemma as a Game with Incomplete Information; in: Journal of Economic Psychology 11, 69 ff.

Campbell, R. and Sowden, L. (ed.) 1985: Paradoxes of Rationality and Cooperation. Prisoner´s Dilemma and Newcomb´s Problem. Vancouver: The University of British Columbia Press.

Cavalli-Sforza, L. anf Feldman, M. 1981: Cultural Transmission and Evolution. Princeton, N.J.: Princeton University Press.

Frank, R. 1987: If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?, in: The American Economic Review Vol. 77/4, 593 ff.

Frank, R. 1988: The Passions within Reason: Prisoner's Dilemmas and the Strategic Role of the Emotions, New York: W. W. Norton.

Gauthier, D. P. 1986: Morals by Agreement. Oxford: Oxford University Press.

Güth, W. 1991: Incomplete Information about Reciprocal Incentives - An Evolutionary Approach to Explaining Cooperative Behavior; Working Paper, University of Frankfurt/M.

Güth, W., and Yaari, M. 1992: An Evolutionary Approach to Explaining Reciprocal Behavior in a Simple Strategic Game; in: U. Witt (ed.), Explaining Process and Change - Approaches to Evolutionary Economics, Ann Arbor: The University of Michigan Press, 23 ff.

Hardin, R. 1982: Exchange Theory on Strategic Basis; in: Social Science Information 2, 251 ff.

Harsanyi, J. C. 1977: Rational Behavior and Bargaining Equilibrium in Games and Social Situations. Cambridge: Cambridge University Press.

Heilbroner, R. 1953/83: The Worldly Philosophers. Harmondsworth: Penguin.

Hirshleifer, J. 1987: The Emotions as Guarantors of Promises and Threats; in: Dupre, J. (ed.), The Latest on the Best. Cambridge: The MIT Press, 307 ff.

Hobbes, Th. 1651/1968: Leviathan. Harmondsworth: Penguin.

Hume, D. 1751/1978: A Treatise of Human Nature. Oxford: Clarendon.

Kant, I. 1798/1991: The Metaphysics of Morals; in: Reiss, H. (ed.), Kant: Political Writings. Cambridge: Cambridge University Press, 131 ff.

Kliemt, H. 1990: The Costs of Organizing Social Cooperation; in: Hechter, M. et al. (ed.), Social Institutions, New York: Aldine-de Gruyter, S. 61 ff.

Kropotkin, P. 1902: Mutual Aid: A Factor in Evolution. New York: Double Day.

Lumsden Ch. J. and Wilson, E. O, 1981: Genes, Mind, and Culture. The Coevolutionary Process. Cambridge: Harvard University Press.

Mackie, J. L. 1985: Morality and the Retributive Emotions; in: J. and P. Mackie (ed.) Persons and Values. Oxford, 206 ff.

Maynard Smith, J. 1964: Group Selection and Kin Selection; in: Nature Vol. 201, 1145 ff.

Maynard Smith, J. and Price, G. 1973: The Logic of Animal Conflicts; in: Nature Vol. 246, 15 ff.

McKelvey, R. D. und Palfrey, Th. 1990: An Experimental Study of the Centipede Game. Social Science Working Paper No 732. Pasadena: California Institute of Technology; to appear in Econometrica 1993.

Parsons, T. 1968: Utilitarianism. Sociological Thought; in: International Encyclopedia of Social Sciences. New York and London.

Raub, W. and Voss, Th. 1986: Conditions for Cooperation in Problematic Social Situations. In: Diekmann, A. und Mitter, P. (ed.): Paradoxical Effects of Social Behavior. Essays in Honor of Anatol Rapoport. Heidelberg und Wien, 85 ff.

Rosenthal, R.W. 1981: Games of Perfect Information, Predatory Pricing and the Chain Store Paradox, in: Journal of Economic Theory, 25, 92 ff.

Smith, A. 1759/1966: The Theory of Moral Sentiments. New York: Kelley.

Posner, R. A. 1981: The Economics of Justice. Cambridge, MA. Harvard University Press.

Selten, R. 1965: Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit, Zeitschrift für die gesamte Staatswissenschaft 121, 301 ff. and 667 ff.

Selten, R. 1975: Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games; in: International Journal of Game Theory, Vol. 4, 25 ff.

Selten, R. 1988: Evolutionary Stability in Extensive Two-Person Games - Correction and Further Development, Mathematical Social Science 16, 223 ff.

Weissing, F. J. 1991: Evolutionary Stability and Dynamic Stability in a Class of Evolutionary Normal Form Games; in: Selten, R. (ed.), Game Equilibrium Models I, Evolution and Game Dynamics, Berlin et al.: Springer, 29 ff.

Discussion Paper Series, CentER, Tilburg University, The Netherlands:

(For previous papers please consult previous discussion papers.)

No.	Author(s)	Title
9201	M. Verbeek and Th. Nijman	Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross Sections
9202	E. Bomhoff	Monetary Policy and Inflation
9203	J. Quiggin and P. Wakker	The Axiomatic Basis of Anticipated Utility; A Clarification
9204	Th. van de Klundert and S. Smulders	Strategies for Growth in a Macroeconomic Setting
9205	E. Siandra	Money and Specialization in Production
9206	W. Härdle	Applied Nonparametric Models
9207	M. Verbeek and Th. Nijman	Incomplete Panels and Selection Bias: A Survey
9208	W. Härdle and A.B. Tsybakov	How Sensitive Are Average Derivatives?
9209	S. Albæk and P.B. Overgaard	Upstream Pricing and Advertising Signal Downstream Demand
9210	M. Cripps and J. Thomas	Reputation and Commitment in Two-Person Repeated Games
9211	S. Albæk	Endogenous Timing in a Game with Incomplete Information
9212	T.J.A. Storcken and P.H.M. Ruys	Extensions of Choice Behaviour
9213	R.M.W.J. Beetsma and F. van der Ploeg	Exchange Rate Bands and Optimal Monetary Accommodation under a Dirty Float
9214	A. van Soest	Discrete Choice Models of Family Labour Supply
9215	W. Güth and K. Ritzberger	On Durable Goods Monopolies and the (Anti-) Coase-Conjecture
9216	A. Simonovits	Indexation of Pensions in Hungary: A Simple Cohort Model
9217	J.-L. Ferreira, I. Gilboa and M. Maschler	Credible Equilibria in Games with Utilities Changing During the Play
9218	P. Borm, H. Keiding, R. Mclean, S. Oortwijn and S. Tijs	The Compromise Value for NTU-Games

No.	Author(s)	Title
9219	J.L. Horowitz and W. Härdle	Testing a Parametric Model against a Semiparametric Alternative
9220	A.L. Bovenberg	Investment-Promoting Policies in Open Economies: The Importance of Intergenerational and International Distributional Effects
9221	S. Smulders and Th. van de Klundert	Monopolistic Competition, Product Variety and Growth: Chamberlin vs. Schumpeter
9222	H. Bester and E. Petrakis	Price Competition and Advertising in Oligopoly
9223	A. van den Nouweland, M. Maschler and S. Tijs	Monotonic Games are Spanning Network Games
9224	H. Suehiro	A "Mistaken Theories" Refinement
9225	H. Suehiro	Robust Selection of Equilibria
9226	D. Friedman	Economically Applicable Evolutionary Games
9227	E. Bomhoff	Four Econometric Fashions and the Kalman Filter Alternative - A Simulation Study
9228	P. Borm, G.-J. Otten and H. Peters	Core Implementation in Modified Strong and Coalition Proof Nash Equilibria
9229	H.G. Bloemen and A. Kapteyn	The Joint Estimation of a Non-Linear Labour Supply Function and a Wage Equation Using Simulated Response Probabilities
9230	R. Beetsma and F. van der Ploeg	Does Inequality Cause Inflation? - The Political Economy of Inflation, Taxation and Government Debt
9231	G. Almekinders and S. Eijffinger	Daily Bundesbank and Federal Reserve Interventions - Do they Affect the Level and Unexpected Volatility of the DM/$-Rate?
9232	F. Vella and M. Verbeek	Estimating the Impact of Endogenous Union Choice on Wages Using Panel Data
9233	P. de Bijl and S. Goyal	Technological Change in Markets with Network Externalities
9234	J. Angrist and G. Imbens	Average Causal Response with Variable Treatment Intensity
9235	L. Meijdam, M. van de Ven and H. Verbon	Strategic Decision Making and the Dynamics of Government Debt
9236	H. Houba and A. de Zeeuw	Strategic Bargaining for the Control of a Dynamic System in State-Space Form

No.	Author(s)	Title
9237	A. Cameron and P. Trivedi	Tests of Independence in Parametric Models: With Applications and Illustrations
9238	J.-S. Pischke	Individual Income, Incomplete Information, and Aggregate Consumption
9239	H. Bloemen	A Model of Labour Supply with Job Offer Restrictions
9240	F. Drost and Th. Nijman	Temporal Aggregation of GARCH Processes
9241	R. Gilles, P. Ruys and J. Shou	Coalition Formation in Large Network Economies
9242	P. Kort	The Effects of Marketable Pollution Permits on the Firm's Optimal Investment Policies
9243	A.L. Bovenberg and F. van der Ploeg	Environmental Policy, Public Finance and the Labour Market in a Second-Best World
9244	W.G. Gale and J.K. Scholz	IRAs and Household Saving
9245	A. Bera and P. Ng	Robust Tests for Heteroskedasticity and Autocorrelation Using Score Function
9246	R.T. Baillie, C.F. Chung and M.A. Tieslau	The Long Memory and Variability of Inflation: A Reappraisal of the Friedman Hypothesis
9247	M.A. Tieslau, P. Schmidt and R.T. Baillie	A Generalized Method of Moments Estimator for Long-Memory Processes
9248	K. Wärneryd	Partisanship as Information
9249	H. Huizinga	The Welfare Effects of Individual Retirement Accounts
9250	H.G. Bloemen	Job Search Theory, Labour Supply and Unemployment Duration
9251	S. Eijffinger and E. Schaling	Central Bank Independence: Searching for the Philosophers' Stone
9252	A.L. Bovenberg and R.A. de Mooij	Environmental Taxation and Labor-Market Distortions
9253	A. Lusardi	Permanent Income, Current Income and Consumption: Evidence from Panel Data
9254	R. Beetsma	Imperfect Credibility of the Band and Risk Premia in the European Monetary System

No.	Author(s)	Title
9301	N. Kahana and S. Nitzan	Credibility and Duration of Political Contests and the Extent of Rent Dissipation
9302	W. Güth and S. Nitzan	Are Moral Objections to Free Riding Evolutionarily Stable?
9303	D. Karotkin and S. Nitzan	Some Peculiarities of Group Decision Making in Teams
9304	A. Lusardi	Euler Equations in Micro Data: Merging Data from Two Samples
9305	W. Güth	A Simple Justification of Quantity Competition and the Cournot-Oligopoly Solution
9306	B. Peleg and S. Tijs	The Consistency Principle For Games in Strategic Form
9307	G. Imbens and A. Lancaster	Case Control Studies with Contaminated Controls
9308	T. Ellingsen and K. Wärneryd	Foreign Direct Investment and the Political Economy of Protection
9309	H. Bester	Price Commitment in Search Markets
9310	T. Callan and A. van Soest	Female Labour Supply in Farm Households: Farm and Off-Farm Participation
9311	M. Pradhan and A. van Soest	Formal and Informal Sector Employment in Urban Areas of Bolivia
9312	Th. Nijman and E. Sentana	Marginalization and Contemporaneous Aggregation in Multivariate GARCH Processes
9313	K. Wärneryd	Communication, Complexity, and Evolutionary Stability
9314	O.P.Attanasio and M. Browning	Consumption over the Life Cycle and over the Business Cycle
9315	F. C. Drost and B. J. M. Werker	A Note on Robinson's Test of Independence
9316	H. Hamers, P. Borm and S. Tijs	On Games Corresponding to Sequencing Situations with Ready Times
9317	W. Güth	On Ultimatum Bargaining Experiments - A Personal Review
9318	M.J.G. van Eijs	On the Determination of the Control Parameters of the Optimal Can-order Policy
9319	S. Hurkens	Multi-sided Pre-play Communication by Burning Money

No.	Author(s)	Title
9320	J.J.G. Lemmen and S.C.W. Eijffinger	The Quantity Approach to Financial Integration: The Feldstein-Horioka Criterion Revisited
9321	A.L. Bovenberg and S. Smulders	Environmental Quality and Pollution-saving Technological Change in a Two-sector Endogenous Growth Model
9322	K.-E. Wärneryd	The Will to Save Money: an Essay on Economic Psychology
9323	D. Talman, Y. Yamamoto and Z. Yang	The $(2^{n+m+1} - 2)$-Ray Algorithm: A New Variable Dimension Simplicial Algorithm For Computing Economic Equilibria on $S^n \times R^m_+$
9324	H. Huizinga	The Financing and Taxation of U.S. Direct Investment Abroad
9325	S.C.W. Eijffinger and E. Schaling	Central Bank Independence: Theory and Evidence
9326	T.C. To	Infant Industry Protection with Learning-by-Doing
9327	J.P.J.F. Scheepens	Bankruptcy Litigation and Optimal Debt Contracts
9328	T.C. To	Tariffs, Rent Extraction and Manipulation of Competition
9329	F. de Jong, T. Nijman and A. Röell	A Comparison of the Cost of Trading French Shares on the Paris Bourse and on SEAQ International
9330	H. Huizinga	The Welfare Effects of Individual Retirement Accounts
9331	H. Huizinga	Time Preference and International Tax Competition
9332	V. Feltkamp, A. Koster, A. van den Nouweland, P. Borm and S. Tijs	Linear Production with Transport of Products, Resources and Technology
9333	B. Lauterbach and U. Ben-Zion	Panic Behavior and the Performance of Circuit Breakers: Empirical Evidence
9334	B. Melenberg and A. van Soest	Semi-parametric Estimation of the Sample Selection Model
9335	A.L. Bovenberg and F. van der Ploeg	Green Policies and Public Finance in a Small Open Economy
9336	E. Schaling	On the Economic Independence of the Central Bank and the Persistence of Inflation
9337	G.-J. Otten	Characterizations of a Game Theoretical Cost Allocation Method
9338	M. Gradstein	Provision of Public Goods With Incomplete Information: Decentralization vs. Central Planning
9339	W. Güth and H. Kliemt	Competition or Co-operation